This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier xx.xxxx/ACCESS.xxxx.xxxxxxx

Applications of Self-Supervised Learning to Biomedical Signals: a Survey

FEDERICO DEL PUP[1,2], and MANFREDO ATZORI[2,3]
[1]Department of Information Engineering, University of Padova, Via Gradenigo 6/b, 35131 Padova, Italy
[2]Department of Neuroscience, University of Padua, Via Belzoni 160, 35121 Padova, Italy
[3]Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais), 2800 Sierre, Switzerland

Corresponding author: Federico Del Pup (e-mail: federico.delpup@studenti.unipd.it).

This work was supported by the EU - Next generationEU program and the Stars@UNIPD - project:MedMax.

ABSTRACT

Over the last decade, deep learning applications in biomedical research have exploded, demonstrating their ability to often outperform previous machine learning approaches in various tasks. However, training deep learning models for biomedical applications requires large amounts of data annotated by experts, whose collection is often time- and cost- prohibitive. Self-Supervised Learning (SSL) has emerged as a prominent solution for such problem, as it allows to learn powerful representations from vast unlabeled data by producing supervisory signals directly from the data. The high amount of recent works employing the self-supervised learning paradigm for the analysis of biomedical signals (biosignals) can make it difficult for researchers to have a complete picture of the current research state. Therefore, this paper aims at outlining and clarifying the state-of-the-art in the domain. The article: briefly summarizes the nature and acquisition modality of the main biosignals; introduces the self-supervised learning method, focusing on the different pretraining strategies; provides a concise overview of the works employing SSL for the analysis of different types of biosignals; provides an overall analysis of critical aspects to consider when employing SSL to biosignals, also highlighting current open challenges. The analysis of the scientific literature highlights the importance of SSL, confirming its potential to improve models' performance and robustness, and to promote the integration of deep learning into clinical tasks.

INDEX TERMS Biosignals, Contrastive Learning (CL), Deep Learning (DL), electrocardiography (ECG), electroencephalography (EEG), electromyography (EMG), multimodal, Self-Supervised Learning (SSL)

I. INTRODUCTION

In the last decade deep learning has emerged as a powerful and versatile tool capable of achieving state-of-the-art performance in various fields. Starting from AlexNet [1], winner of the 2012 Imagenet Large Scale Visual Recognition Challenge (ILSVRC) [2], many of the biggest companies have invested lots of resources to promote and introduce deep learning applications in their products and software. Notorious examples are Google DeepMind's AlphaZero [3], a reinforcement learning algorithm capable of winning against the strongest humans and computer engines on various board games (e.g., chess, go, shoji), Google DeepMind's AlphaFold [4], winner of the 13th and 14th Critical Assessment of Techniques for Protein Structure Prediction (CASP), and the novel OpenAI's ChatGPT[1], considered a fundamental step in Natural Language Processing (NLP). The mentioned examples demonstrate how deep learning can be successfully applied in various research area; therefore, medicine was not excluded by the "golden fever" of Artificial Intelligence (AI). Looking at PubMed[2], one of the most used search engines for biomedical literature [5], it is possible to see that the number of yearly published works involving deep learning has increased from less than 300 in 2016 to approximately 17 000 in 2022 (a remarkable increase of approximately 5 700%). However, despite the rocketing number of applications, the use of deep learning is still limited in common clinical

[1][Online]. Available: https://openai.com/blog/chatgpt/
[2][Online]. Available: https://pubmed.ncbi.nlm.nih.gov/

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

practice [6]. Deep neural networks are usually trained in a supervised way, where a manually labeled dataset is fed to train, optimize and test the model. At first, given the novelty of the field, this approach was able to often outperform previous state-of-the-art algorithms based on more naive approaches [7]. More recently, considering the increasing complexity of models and tasks where deep learning can be involved, limitations have started to be highlighted [8]. Training large neural networks that can generalize well in the biomedical domain requires a huge amount of highly heterogeneous annotated data, which is difficult to collect in medical research [9]. In fact, manually labeling medical data is a time-consuming task that only experts in the field can do. Furthermore, their collection is often hindered by ethical (e.g., trial approval, anonymization) and economical aspects, which make data provision and annotation an extremely challenging task. In contrast, thanks to the digitization of the healthcare sector, a large amount of unlabeled data is generated every day, with an order of magnitude already reaching the exa-scale [10]. Exploiting them could greatly improve the performance and robustness of deep learning models, which is why the research community has started to propose novel unsupervised solutions.

Self-Supervised Learning (SSL) has emerged as one of the most prominent paradigms in this context. Its goal is to learn robust general-purpose representations from the data by exploiting an auxiliary task (pretext task); then, transfer the acquired knowledge to a new model designed to solve the target (medical) task. Self-supervised learning has been successfully applied in many fields, such as natural language processing [11], computer vision [12], speech recognition [13], and robotics [14]. In medical research, computer vision is the most investigated area [15]. Here, self-supervised learning is employed for classification, segmentation, registration and reconstruction of different types of images, from 2D microscopy for digital pathology [16] to 3D MRI (magnetic resonance imaging) [17]. The interested reader can consult the work of Saeed *et al.* [18] and that of Xu [19], who have already reviewed SSL implementations in the medical imaging domain.

Biomedical signals (biosignals) represent a fundamental resource in the medical domain, including many modalities such as electroencephalography (EEG), electromyography (EMG), and electrocardiography (ECG). Moreover, with the progress of the IoT (Internet of Things) and the spread of wearable devices, their role is increasingly becoming more relevant, especially in telehealth and precision medicine [20]. As a matter of fact, several researchers have already proposed SSL strategies for the analysis of biosignals. However, considering the large and constantly growing number of publications, it is difficult to keep up with the progress of the state of the art. A review targeting SSL applications to biosignals is not available according to the best of our knowledge. In fact, previously cited works focus on different types of data (medical imaging) [19], [20], specific biomedical signals (EEG) [21], or specific self-supervised learning paradigms (contrastive learning) [15]. Moreover, they often tend to extensively describe SSL pretraining strategies and the surveyed works but do not put the same effort into discussing special aspects to consider when employing existing SSL techniques for a specific biosignal analysis task (the work of Rafiei *et al.* [21] for EEG data is an exception), which are crucial for effectively designing novel strategies. Therefore, this paper aims at solving these limitations by providing a resource where readers can receive an outline of the main principles behind the most commonly used SSL frameworks for the analysis of biomedical signals and have a overview of the current state-of-the-art of the domain, regardless of the nature of the signal or of the investigated self-supervised paradigm.

The rest of the work is organized as follows. Section II provides a brief description of the most important types of biosignals, with a focus on the ones encountered during the survey. Sections III and IV introduce the self-supervised learning paradigm, describing its main concepts and different pretext task strategies. In Section V, a brief description of the survey methodology is provided to the reader. Section VI reports and analyzes SSL applications for the analysis of different types of biosignals (e.g., ECG, EEG, and EMG), also considering multimodal approaches. Section VII aims to answer to different questions related to the application of SSL for biosignals analysis, while also providing a description of critical issues and open challenges. Finally, Section VIII summarizes the most important outcomes of the work.

II. BIOSIGNALS

As per Bansal's Real-Time Data Acquisition in Human Physiology [25]: "Biological signals, or Biosignals, are space, time, or space–time records of a biological event such as a beating heart or a contracting muscle. The electrical, chemical, and mechanical activity that occurs during these biological events often produces signals that can be measured and analyzed. Biosignals, therefore, contain useful information that can be used to understand the underlying physiological mechanisms of a specific biological event or system, and which may be useful for medical diagnosis".

Most of the biosignals are of the electrical type, collected by electrodes placed in specific parts of the body (e.g., head for electroencephalography, chest and limbs for electrocardiography, eyes' region for electrooculography), generally in a noninvasive way. Moreover, with the spread of wearable devices, the acquisition and collection of various types of biological signals has become much easier, hence their exploitation for clinical tasks [26]. For example, Continuous Glucose Monitoring (CGM)

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

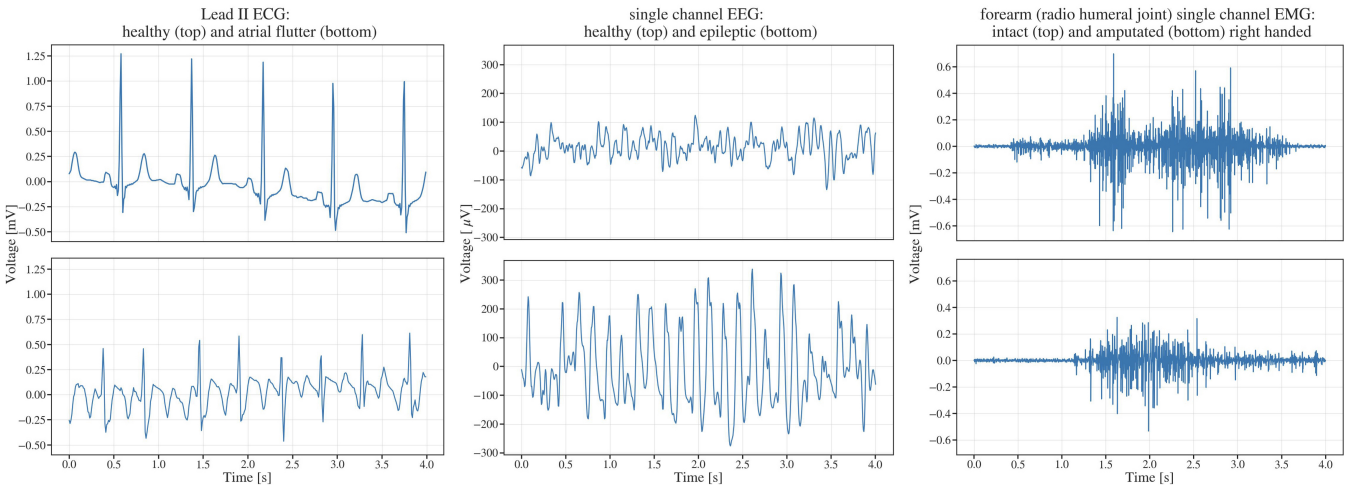

FIGURE 1. **Example of four seconds of three different biosignals (ECG, EEG, and EMG) with normal (top) and abnormal (bottom) conditions. Left: lead II ECGs selected from the PTB-XL dataset [22]. Healthy subject is subject 18, and pathological subject (atrial flutter) is subject 33. Middle: single-channel EEGs selected from the BONN EEG dataset [23]. Healthy subject is subject 2 from set B, while pathological (epileptic) subject is subject 6 from set E. Right: single channel EMGs selected from the NinaPro dataset [24]. Intact right-handed subject is subject 16 from dataset 2, while amputated right-handed subject is subject 4 from dataset 3. Note how, regardless of their type, it is possible to spot some differences in amplitude and/or waveforms between normal and abnormal biosignals.**

devices can help diabetic people manage their disease by detecting real-time variations of the blood glucose concentration at intervals of usually one, three, or five minutes [27]. Moreover, other devices like smart wristbands (e.g., Empatica© E4) can simultaneously record different types of biosignals by means of multiple sensors, introducing the possibility to combine their acquisitions with other types of data to improve the diagnosis and prognosis of several pathology. Biosignals collected by wearable devices may include blood volume pressure, electrodermal activity (i.e., variation in the electrical properties of the skin), temperature readings, motion-based activity data, and many more. Give a complete description of all the biosignals is beyond the scope of this work. Nevertheless, it is important to at least introduce the main ones encountered during the survey:

- **Electrocardiography (ECG)**: these types of signals record the electrical activity of the heart. ECGs are generally recorded with the 12-lead method, which consists of placing ten electrodes, six on the chest and the remaining on the limbs, to calculate a set of electric potentials. The combination of the measurements from all the electrodes gives a unique quantitative and spatial information about the heart's electrical activity, called lead. An ECG machine processes the information coming from all 12 leads to produce a graphical representation. ECGs possess a particular structure (P wave, QRS complex, ST segment, T wave, and U wave) given by the sequential repolarization and depolarization of the heart's atria and ventricles. Unusual variations in the amplitude, time, or frequency of these structures provide information about the normal or abnormal activity of the heart, thus leading to the diagnosis of a particular pathology [28] (see exemplary ECGs provided in the left part of figure 1);
- **Electroencephalography (EEG)**: these types of signals record the electrical activity of the brain cells generated by the exchange of ions between the inside and outside of the neurons. EEGs are usually recorded by placing several electrodes around the subject's scalp in specific configurations, which can vary depending on the number of electrodes and the study objective. EEG signals are really complex and are usually analyzed both in the time and frequency domains. In fact, clinically relevant information for diagnostic and prognostic purposes can be retrieved by looking at specific bandwidths of the signal, namely: delta (0.3–4 Hz), theta (4–8 Hz), alpha (8–14 Hz), beta (14–30 Hz), and gamma (>30 Hz). EEGs are widely adopted by neuroscientists for cognitive tasks as well as for the study of several neurological disorders such as epilepsy (exemplary EEGs provided in the middle part of figure 1), dyslexia, and mental diseases [29];
- **Electromyography (EMG)**: these types of signals record the electric currents that are generated during muscle contraction. EMGs are usually recorded by surface electrodes, but more invasive types like needle electrodes can be adopted to improve the signal-to-noise ratio and to get access to single motor unit action potentials (MUAP). EMGs are generally used to detect anomalies in the activity of the muscles (e.g., myopathy, neuropathy) as well as in biomechanics for the development of body prosthetics [30] (see exemplary EMGs provided in the right part of figure 1);

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

- **Other types of biosignal**: other biosignals used for various clinical tasks and therefore worthy of being mentioned are the *magnetoencephalography* (MEG), which measures the magnetic field generated by the activity of brain cells and has many applications such as brain connectivity, cognitive studies on newborns, and epilepsy research; the *phonocardiography* (PCG), which measures the sound produced by the heart's beat and is used for the detection of heart diseases; the *electroretinography* (ERG), which measures the electrical activity of various cell types in the retina and is mainly used for diagnostic reasons; the *electrooculography* (EOG), which measures the electric potential that is generated by the cornea and the retinal activity during eye movement; eye tracking data, which measures the orientation of the eye in space or the position of the eye with respect to the subject's head [31]. Unfortunately, researchers have not yet used most of these signals to train deep learning models in a self-supervised way. However, future works may include them, especially in multimodal approaches.

III. SELF-SUPERVISED LEARNING

Training deep neural networks with fully supervised methods requires large amounts of data. In medical research, however, it is usually difficult to assemble very large datasets. The acquisition of medical data is in fact expensive in terms of time, costs and administrative procedures (e.g., ethics). It also requires specific instrumentation and human volunteers. Moreover, data annotation can be performed only by medical experts in a laborious and time-consuming process. Ultimately, medical data are highly heterogeneous (e.g., instrumentation, acquisition protocols and settings, subject-variability), and the model's robustness and generalization capability are inherently affected by that [32].

In contrast, the amount of unlabeled data is enormous. For this reason, researchers have started to investigate new methodologies to exploit unlabeled data [33] such as semi-supervised learning [34], weakly-supervised learning [35], or self-supervised learning, as described in this section.

Self-supervised learning attempts to address the issue of having limited annotated data by extracting general-purpose features from vast unlabeled data [36]; hence, it is usually referred to as an unsupervised technique. Despite that, self-supervised learning differs from common unsupervised methods like clustering [37] or Principal Component Analysis (PCA) [38]. In particular, clustering techniques aim at finding groups of similar objects by agglomerating or separating samples based on specific metrics (distance functions) designed to evaluate the grade of dissimilarity between the investigated data. PCA is instead used to reduce the dimensionality of a dataset by finding new variables that are linear functions of the original ones, that successively maximize variance, and that are uncorrelated with each other [39]. Both techniques are mainly used as exploratory tools for data analysis with the goal of inferring statistical properties of the investigated feature set. Moreover, they don't include any type of label, nor they are used to predict some outcome from unobserved data, like in supervised approaches. On the contrary, SSL aims at predicting part of its input from other parts of its input, converting the unsupervised problem into a supervised one (hence its name). As it will be clarified in the next section, self-supervised learning can in fact generate its own form of supervision directly from the data; hence, it can use way more supervisory signals than standard fully supervised approaches. That's why it is more proper and less misleading to allocate SSL algorithms in a separate category rather than trying to associate them with other unsupervised methods.

Figure 2 summarizes how the self-supervised learning paradigm works. First, a deep neural network is trained to solve an auxiliary task, also called *pretext task*, *upstream task*, or simply *pretask*, whose primary goal is to learn general-purpose features of the given data without having access to any sort of external supervision. During this phase, no information about the target (medical) task or the real (physiological) meaning of the given data is explicitly used. Moreover, no interest is given to the model's performance, as the pretext task has (often) no connection to the target one, and it is designed with the assumption that solving it requires the network to learn useful information intrinsic to the data; in other words, model them. Although pretraining strategies can highly differ from each other, this phase usually includes the generation of artificially created pseudo-labels from the unlabeled dataset, here used as the target variable. Training samples are then fed to the model in order to predict the constructed target. Finally, model predictions are used to calculate the value of a given objective function, which is then used to update the model weights with backpropagation. Once the model is pretrained, the weights of its feature extractor (encoder) are transferred to a new model, which will be trained to solve the target task, usually called *downstream task*. The new model shares the same backbone structure, while its head (final set of hidden layers) is slightly modified to make it compatible for the downstream task, for example by adding a softmax or a regression layer in case of classification or regression problems, respectively. Model transfer is performed by applying transfer learning, a method that consists of employing the knowledge that has been learnt in a source task (here upstream task) to another target task (here downstream task) in order to improve the performance and generalization capability of the new model [40].

The final step, which is performed after the encoder's

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

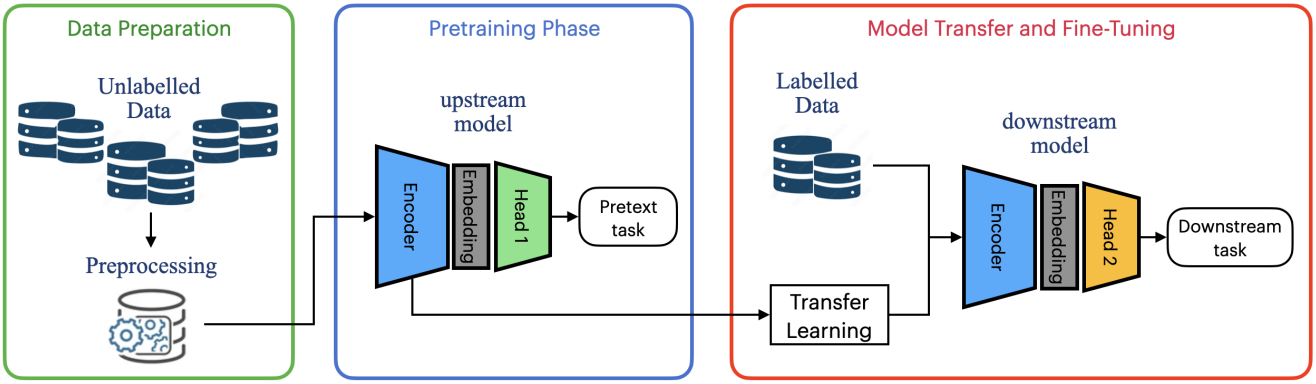

FIGURE 2. **A simple schematic representation of the self-supervised Learning paradigm. First, a model is pretrained with only unlabeled data to solve an auxiliary task (pretext task). Then, the backbone's weights are transferred to the downstream model, which is then fine-tuned with the limited amount of labeled data.**

weights are transferred to the new downstream model, consists of learning more task-specific features using the limited amount of labeled data in a process called fine-tuning. The fine-tuning phase shares many similarities with a standard fully supervised training procedure; the main difference resides in the fact that the model weights, instead of being randomly initialized, originated from the solved pretext task. Another important difference is that, as described in [21], it is common practice to divide the fine-tuning process into two steps. The first consists of freezing all the backbone weights and updating only the modified final hidden layers; then, conclude the training process with the whole network unfrozen.

In conclusion, self-supervised learning, although more complex than a standard fully supervised approach, is supposed to help improve accuracy and mitigate overfitting in contexts where the amount of labeled data is limited or where multiple heterogeneous datasets can be aggregated, which is likely the biomedical context.

IV. PRETEXT TASKS

Pretext tasks are the core of the self-supervised learning paradigm. Although they are designed for the same scope, which is learn general-purpose feature without having access to manual annotated data, pretext tasks can accomplish their goal in different ways. Some of them have been developed for specific types of data, such as the Rubik's cube method for 3D images (e.g., magnetic resonance imaging) [41]. Others are more versatile and allow researchers to work with different types of data. This section focuses on approaches that are compatible with biosignals and that were encountered during the survey. Various classification schemes have been proposed to organize the pretext tasks, depending on the domain of application [42]. Here, methodologies will be grouped into the following three categories: predictive, generative, and contrastive learning pretext tasks.

A. PREDICTIVE PRETEXT

Predictive pretext is a family of supervised pretraining methods characterized by the construction of classification or regression problems as an auxiliary task. This approach makes use of artificially created pseudo-labels, which are assigned to the unlabeled data, to pretrain the model in a supervised way. The generation of the pseudo-labels, which needs to be automatic and knowledge-free, is what really differentiates one approach from the other. For example, one can simply construct a transformation recognition problem, where single or multiple transformations (e.g., scaling, permutation, time shift, noise addition) are applied to the original sample with the goal of predicting or classifying them. Others can exploit specific (biological) properties of the signal and construct more complex targets to predict [43]. An example of such a strategy can be found in [44]. Here, the authors have applied two different sets of transformations to EEG data with the goal of producing abnormal samples. The first transformation amplifies portions of the signal in the time domain, while the second alters the original sample in the frequency domain. Both original and transformed samples were fed to the pretraining model in order to predict the type of transformation, thus building a 3-class classification problem. The model was pretrained to optimize the cross-entropy loss, and its head (the final softmax layer) was discarded during model transfer. Similar protocols can be applied to other biosignals or to build regression predictive pretext tasks. For example, authors in [45] have built a regression task based on the prediction of features extracted directly from the ECG signal (characteristic intervals and amplitudes). Predictive pretext tasks are fairly easy to implement and do not require many computational resources compared to other methods. However, the specificity of the task has a strong impact in the quality

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

of representations. Therefore, careful consideration must be given to its design, as wrong choices could degrade model performance.

B. GENERATIVE PRETEXT

Generative pretext [46] is a family of unsupervised methods widely used in natural language processing (like BERT [47]) which is living a new life in other domains like computer vision and signal processing [48]. Its goal is to train general-purpose features by learning either to regenerate an augmented version of the input data or to generate new samples from the same distribution of the training repository. Since the pretext task is treated as a generative problem, architectures like auto-encoders [49] or Generative Adversarial Networks (GAN) [50] are utilized in this category. In the signal domain, the most adopted generative pretext task is masked modeling, whose goal is to learn robust representations by reconstructing a portion of the signal that was previously cropped or masked. Masked modeling is widely adopted for other types of data as well. Two examples are the masked autoencoders for imaging data [51] and the work presented in [52] for audio data. Another example of such a strategy can be found in [53]. Here, authors have applied a set of transformations to EEG data in order to generate new corrupted samples. Such transformations not only include the cited masking operation but also other ones typically employed in predictive pretext task such as the noise addition, the moving average filtering, or the EEG channel dropout. However, in contrast to predictive pretraining strategies, no artificial pseudo-labels were generated, and the original samples were used directly as the predictive target. In this setting, the Mean Square Error (MSE) between the output of the model and the original sample was used as the objective function to evaluate the quality of the reconstructed samples and update the model weights. Practical challenges associated with generative pretext, such as the higher computational costs and repository size required to efficiently pretrain the model, make this approach rarely adopted compared to other supervised pretext tasks. In fact, GAN-based approaches like the one proposed in [54] require learning two different neural blocks: a generator, responsible for creating new samples, and a discriminator, responsible for distinguishing between the original and the generated sample, which is the only block that is kept after pretraining. The presence of two different neural blocks, usually with numerous parameters, inevitably increases the computational demand and, consequently, the training time and the needed GPU memory.

C. CONTRASTIVE LEARNING PRETEXT

Contrastive Learning (CL) is a family of methods that aims at learning robust general-purpose representations from the data by embedding augmented versions of the same sample close to each other while trying to push away representations from different samples [55]. This goal is achieved either by learning to discriminate between similar (positives) and dissimilar (negative) samples, or by maximizing only the agreement between pairs of similar views. Data augmentation is the core of contrastive learning. Positive and negative samples are generated by applying a set of transformations to the original sample (e.g., noise addition, scaling, permutation, horizontal or vertical flip), which aim at introducing some differences while at the same time preserving the data global features. Contrastive learning has gained enormous attention due to its simplicity and effectiveness in training general-purpose encoders. For this reason, a large variety of approaches can be found in the literature, usually employing siamese architectures (weight-sharing neural networks applied on two or more inputs) [56] to compare the augmented samples. Here are reported only those baseline methodologies that have been applied in works selected during the survey, whose schematic views are collected in Figure 3:

(a) **CPC (2019)**: Contrastive Predictive Coding (CPC) is a modality-agnostic framework designed to suit any type of data (e.g., images, text, speech, signals) [57]. Its goal is to predict high-level information of future time steps of a sample given a series of past ones. However, instead of simply trying to predict future observations, CPC aims to learn the underlying shared information between different parts of the (high-dimensional) signal. Figure 3(a) summarizes how CPC works. First, sequences of observations x_{t+k}, $k \in \mathbb{Z}$, are passed to a non linear encoder to produce a set of latent representations z_{t+k}; then, latent representations of the past portion of the signal are fed into an autoregressive model, which is used to summarize all the encoded information and produce a context latent representation c_t. Finally, the context latent representation is used to predict the latent representation of future portions of the signal (target). The encoder and the autoregressive model are trained to jointly optimize a loss based on noise-contrastive estimation (NCE) [58], which is called InfoNCE loss.

(b) **SimCLR (2020)**: A simple framework for Contrastive Learning Visual Representation (SimCLR) is an end-to-end framework designed to learn high-quality representations by maximizing the agreement between differently augmented views of the same data example via a contrastive loss in the latent space [59]. SimCLR relies on two simple key ideas. The first is to use heavy random data augmentation; the second is to adopt large batch sizes rich of negative examples. Figure 3(b) illustrates how SimCLR works. Each sample x is

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

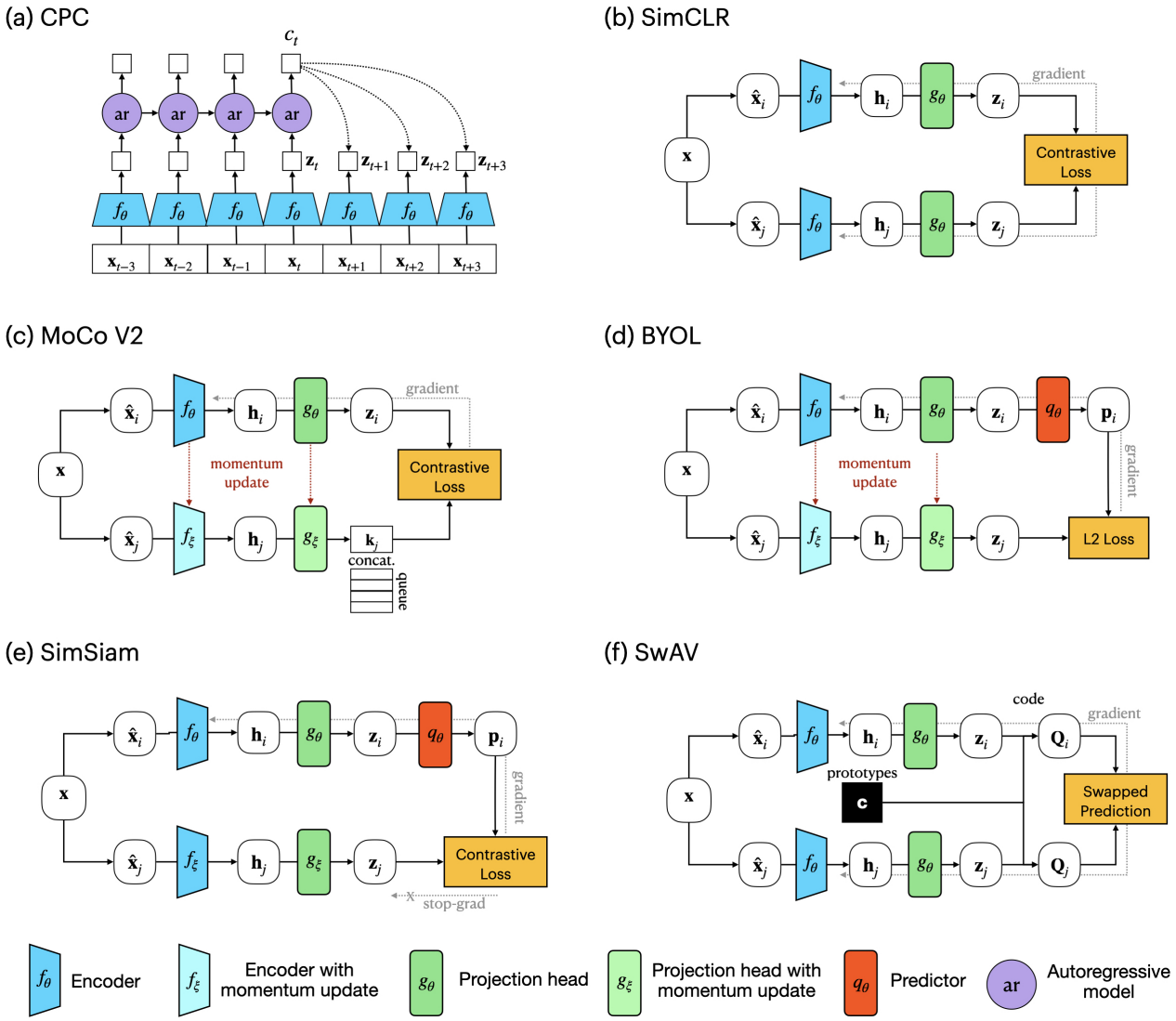

FIGURE 3. **Schematic view of some contrastive learning frameworks. (a) Contrastive Predictive Coding (CPC); (b) Simple Contrastive Learning (SimCLR); (c) Momentum Contrast (MoCo); (d) Bootstrap Your Own Latency (BYOL); (e) Simple Siamese (SimSiam); (f) Swapping assignment between views (SwAV). x denote the original sample, x̂ its augmented version, h the encoder output, z the latent representation, k the new enqueued keys in MoCo, p the prediction of the online network in BYOL, $\mathbf{c}_t$ the context latent representation in CPC. Momentum network modules are represented with a lighter color compared to their online counterparts in order to highlight their little difference in weight values. Also note that all methods are similar between to other but have a clearly distinct peculiarities.**

augmented twice with randomly selected transformation functions. Then, each of the augmented samples is fed to a backbone encoder to produce a set of representations $\mathbf{h}$; after that, representations are passed to a small neural block called projector head, which will output a set of projections $\mathbf{z}$ in a new latent space. Finally, projections are used to maximize the agreement between positive pairs, i.e., pairs of augmented samples of the same original data. The encoder and projector head are trained to jointly optimize the normalized temperature-scaled cross entropy loss (NT-Xent), defined as:

$$\mathcal{L}_{\mathrm{i,j}} = -\log \frac{\exp(\mathrm{sim}(\mathbf{z}_\mathrm{i}, \mathbf{z}_\mathrm{j})/\tau)}{\sum_{\mathrm{k}=1}^{2\mathrm{N}} \mathbb{1}_{[\mathrm{k}\neq\mathrm{i}]} \exp(\mathrm{sim}(\mathbf{z}_\mathrm{i}, \mathbf{z}_\mathrm{k})/\tau)} \quad (1)$$

with $\mathrm{sim}(\mathbf{z}_\mathrm{i}, \mathbf{z}_\mathrm{j})$ the cosine similarity between two projections, and τ the temperature parameter.

(c) **MoCo (2020)**: Momentum Contrast (MoCo) [60] is a method that, in its updated version (MoCo V2 [61]), outperformed end-to-end frameworks like SimCLR. MoCo took the problem of learning good representations by performing look-up operations on a large dictionary rich of negative examples, which is continuously updated to keep it consistent

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

during training. The dictionary, which can be considered an improvement of the memory bank introduced in [62], allows for lessening the memory burden while keeping the number of negative pairs sufficiently high. In fact, the dictionary size can be much larger than a typical batch size and is treated as a queue, where newer keys from the current mini-batch are enqueued while the oldest are removed. Unlike SimCLR, where the two branches of the siamese architecture share the same parameters, MoCo adopts two different networks identical in structure (encoder plus projection head) but different in weight values (see Figure 3(c)). The first is the online network (parametrized by $\boldsymbol{\theta}$), which is responsible for generating a set of projections $\mathbf{z}$ (as in SimCLR). The second is the momentum network (parametrized by $\boldsymbol{\xi}$), which is responsible for encoding the new dictionary keys $\mathbf{k}$ to be enqueued. The online network is trained to optimize the InfoNCE loss and is updated through stochastic gradient descent. On the contrary, since the dictionary does not allow back-propagation on the momentum network, the latter is updated with an exponential moving average of the online network weights, defined as:

$$\boldsymbol{\xi} = \mathrm{m}\boldsymbol{\xi} + (1 - \mathrm{m})\boldsymbol{\theta} \tag{2}$$

with $\mathrm{m} \in [0, 1)$ momentum coefficient, usually bigger than 0.995.

(d) **BYOL (2020)**: Bootstrap Your Own Latent (BYOL) is a method that, unlike SimCLR or MoCo, uses neither negative pairs nor contrastive losses [63]. In particular, BYOL sets up a regression task as the learning problem, where the embedding of one augmented version of a sample is used to predict the embedding of another augmented version of the same data. It is important to note that using only positive pairs could potentially lead to a collapsing solution [64], i.e., the trend of siamese architectures to "collapse" to a constant output. However, authors empirically demonstrated that an asymmetrical architecture and the momentum encoder could avoid this problem. In fact, as can be seen in Figure 3(d), BYOL adopts two different networks to learn. The first is the online network (parametrized by $\boldsymbol{\theta}$), which added a predictor block q_θ after the usual encoder and projector blocks, and is used to make the predictions. The second is the target network (parametrized by $\boldsymbol{\xi}$), which is used to provide the regression target to be predicted by the online network. During training, two augmented versions of a sample are fed into the networks. Then, outputs are ℓ_2-normalized and the mean square error (MSE) is calculated. Finally, the online network is updated through stochastic gradient descent, while momentum updates are used to change the target network weights.

(e) **SimSiam (2020)**: Simple Siamese Representation Learning can be considered a simplified version of BYOL without the momentum encoder [65]. The key element of this minimalist approach is the stop-grad operation. The authors empirically showed that this operation is sufficient to avoid the collapsing solution and no momentum encoder, like in BYOL, is necessary. However, the gain in simplicity is counterbalanced by a slight drop in performance.

(f) **SwAV (2020)**: Swapping Assignments Between Views (SwAV) is a cluster-discrimination-based framework [66]. SwAV does not directly compare features extracted from different transformations of the same sample, like in previous approaches. Instead, it combines online clustering with a swapped prediction mechanism to enforce consistency mapping between augmentations of the same original data. Figure 3(f) illustrates the structure of SwAV. In particular, each augmented sample is fed into an encoder to produce a vector representation. Representations are then ℓ_2-normalized and mapped to a set of trainable normal vectors, i.e., prototypes, thus computing a "code" $\mathbf{Q}_\mathrm{i}$. In other words, prototypes can be considered as the clusters where the data are being partitioned and codes the results of the online clustering. Finally, with the assumption that different views of the same image should maintain similar information, the model is trained to predict the cluster assignment of a view from the representation of another view (swapping prediction).

Aside from the ones already listed, other contrastive learning methods can be employed with time-series data. Such methods mainly differ from the structure of the network, the formulation of the contrastive loss, and the way negative samples are exploited. Few examples are PIRL [67], Barlow Twins [68], VICReg [69], W-MSE [70], TNC [71], MoCo V3 [72], and DINO [73]. The interested reader can consult the work of Balestriero et al. [74], which provides a more detailed analysis of the self-supervised learning paradigm, with lots of insights on critical aspects of its implementation.

D. NOVEL METHODS FOR TIME-SERIES DATA

The previously cited baseline methods were used as a reference for the development of novel approaches for time-series data that were not specifically designed for medical applications, but still tested on medical repositories. For example, Cheng *et al.* [75] proposed a subject-aware contrastive learning method for biosignals whose core element was the addition of an adversarial subject identifier module to promote subject-invariance during pretraining and mitigate the nega-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

tive effects of inter-subject variability. Gorade *et al.* [76] proposed a BYOL-based approach based on the combination of two different sets of projector plus predictor designed to extract, respectively, low- and high-frequency characteristic features from the embedding. Zhang *et al.* [77] developed a contrastive pretraining method which promoted the alignment of time- and frequency-based representations projected in a shared latent space. Ultimately, Wickstrøm *et al.* [78] proposed a novel contrastive learning approach that combined a custom contrastive loss with a new data augmentation scheme designed to generate new data by mixing two training samples. All methods listed in this subsection demonstrate that ideas from other research areas can be successfully imported into the medical domain. However, as it will be discussed in section VII, special considerations about the physiological nature of the signal and the target clinical task must be taken into account in order to avoid failures in the application of SSL strategies.

V. SURVEY METHODOLOGY

This section summarizes the methodology followed to search and identify relevant literature on self-supervised learning for the analysis of biosignals. To summarize, it consists of a first selection of papers from various literature sources, followed by multiple exclusions, if necessary, using specific criteria. For the literature search, the following bibliographic databases were used as primary references:

- PubMed[3]
- IEEE Xplore[4],
- Springer Link[5],
- ScienceDirect[6],

In addition, the research was extended to other sources of literature, namely:

- Google Scholar[7]
- ArXiv Preprints[8]

Google scholar allows researchers to automatically gather articles from the previously listed literature databases. However, we preferred to investigate directly all the single sources and use scholar only to double-check and refine the research in case of possible missing works. Moreover, we carefully checked arXiv preprints before considering their inclusion, as they did not undergo a full peer-review process.

Each source was queried by combining a set of selected keywords, using at first only general terms (e.g., self-supervised learning, contrastive learning, time-series, biosignal); then, research was refined by adding more specific terms related to each type of biosignal (e.g., ECG, EEG, EMG, EOG). An example of such an approach, using the Google Scholar query format for simplicity, is reported below:

1) "Self-Supervised Learning" AND "time-series";
2) "Self-Supervised Learning" AND "biomedical";
3) "Self-Supervised Learning" AND "biosignals";
4) "Self-Supervised Learning" AND "wearable sensors";
5) "Self-Supervised|Contrastive Learning" AND "Electrocardiogram|ECG";
6) "Self-Supervised|Contrastive Learning" AND "Electroencephalography|EEG";
7) "Self-Supervised|Contrastive Learning" AND "Electromyography|EMG";
8) "Self-Supervised|Contrastive Learning" AND "Electrooculogram|EOG".

Self-supervised learning is a novel technique that has only recently made its way into medical research. In addition, this field is very mutable and the state of the art can rapidly change. For this reason, only works published no earlier than 2016 were considered, focusing on the period 2019 - 2023, when their number has increased considerably. In particular, we included only those papers that adopted self-supervised learning on biosignals to solve medical tasks. We also considered publications that present novel SSL methods not specifically designed for medical tasks but still tested on biomedical datasets, gathered for organizational reasons in the subsection IV-D. From the selected list, we excluded works that adopted the same SSL methodology to solve a particular task or works that have been updated by another one. In those cases, we kept the one that we considered the most relevant by weighting several factors such as the number of citations, the impact factor of the journal or conference, and the type of work. Finally, we considered research works cited in the bibliography or in the related works sections of the selected papers.

VI. SELF-SUPERVISED LEARNING ON BIOSIGNALS

The survey resulted in a selection of 61 works describing SSL applications for the analysis of biosignals. As can be seen in Figure 4, there is a high imbalance between applications on ECG or EEG signals and other types of data, probably associated with the higher availability of public datasets. Taking this into account, the results were grouped into four categories, namely: SSL on ECG, SSL on EEG, SSL on other types of biosignals, multimodal SSL with biosignals. Each category will present the investigated medical tasks and the adopted pretraining strategies, delving into those works that present novel SSL approaches. At the end of each section, a summary table reports a synthesis of the main information for each of the presented works, namely: upstream task, downstream task, datasets used, year of release, and, if necessary, the type of data. Tables were sorted by year

[3][Online]. Available: https://pubmed.ncbi.nlm.nih.gov/
[4][Online]. Available: https://ieeexplore.ieee.org/
[5][Online]. Available: https://link.springer.com/
[6][Online]. Available: https://www.sciencedirect.com/
[7][Online]. Available: https://scholar.google.com/
[8][Online]. Available: https://arxiv.org/

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

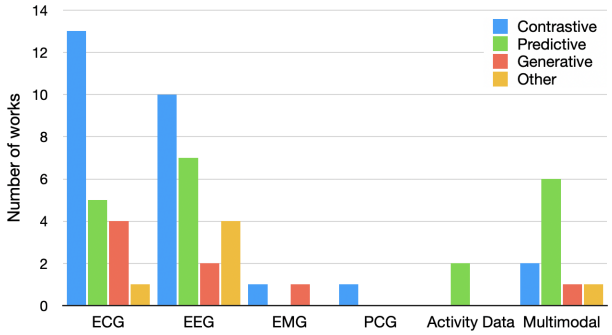

FIGURE 4. Number of works per SSL strategy grouped by the type of biosignal (ECG: electrocardiography, EEG: electroencephalography, EMG: electromyography, PCG: phonocardiography) adopted and the type of upstream task. The "other" category refers to those works that have tested multiple SSL pretraining strategies or have proposed hybrid approaches, i.e., a combination of the previous three.

and author names. Moreover, works sharing the same downstream task were grouped together to improve the organization and consultation of big tables.

A. SELF-SUPERVISED LEARNING ON ECG

ECG is one of the two major categories of biosignals where self-supervised learning has been adopted until now. Out of all the investigated tasks, classification of cardiac pathology (e.g., arrhythmia) plays a central role in SSL ECG-based analysis, with 19 out of 23 works evaluating such a medical task. This aspect reflects the high demand for integrating deep learning models into decision support systems to be used in real-world scenarios, which still suffer from the great variability associated with such data. In fact, most of the datasets listed in table 1 are open repositories released by large hospitals or collected for highly competitive challenges (CinC datasets) organized to improve ECG medical analysis.

In contrast to the identified trend in the investigated downstream tasks, the choice of the pretraining strategy is highly variable but reveals an overall preference for contrastive learning pretext tasks (see table 1). Moreover, the provided survey reveals that works often attempted to exploit biological properties of the ECG signal during pretraining, for example by exploiting its peculiar waveform [45], its periodicity [81], [84], or its associated variability [90].

1) Cardiac Pathology

Concerning single- or multi-class pathology classification tasks, contrastive learning was the primary choice in most of the studies. Nakamoto *et al.* [110] adapted the baseline MoCo for left ventricular systolic dysfunction detection, while Lai *et al.* [115] improved the loss of the same algorithm to recognize 60 diagnostic terms on a large scale private Dataset. Mehari *et al.* [109] compared baseline contrastive learning approaches (e.g., SimCLR, BYOL, SwAV, CPC) to assess their ability to extract good representations from the ECG signal, while Soltanieh *et al.* [111] provides an extensive analysis on the efficacy of different data augmentations. Lee *et al.* [108] proposed a variant of the contrastive learning algorithm VICReg (VIbCReg) which slightly modified the loss function and included, after the projector of the siamese network, an additional iterative normalization layer. Gopal *et al.* [107] leveraged the unique spatiotemporal properties of the ECG signal by adopting a physiologically 3D augmentation technique to generate the positive pairs for the contrastive learning pretraining phase. Liu *et al.* [113] proposed a joint cross-dimensional contrastive learning method that consist of pretraining the model to maximize the similarity between positive pairs of ECG signals as well as between the ECG and its 2-D image representation. It is important to note that contrastive learning was not a unilateral choice, and other pretext tasks were investigated this type of problem. In particular, Yang *et al.* [112] and Gedon et al. [105] adopted masked modeling for the representation learning part of the model, both achieving comparable or slightly superior results to fully supervised training.

Given the large amount of supervision that can be provided from some of the available free open datasets (e.g., PTB-XL [22]), results were not always superior to fully supervised state-of-the-art methods but still comparable. For example, Liu *et al.* [113] reported an absolute drop in accuracy and F1-macro score of respectively 0.012 and 0.033 on the PTB-XL dataset, with similar results on the CPSC2018 dataset [79]. However, comparable performances were achieved using only half the available labels, showing the robustness of SSL strategies against drops in the label ratio. This demonstrated that self-supervision has the potential to improve the learning process but requires further advances before becoming the new golden standard for those challenging problems.

2) Arrhythmia classification

Upon the pathology investigated, arrhythmia seems to be the most common use case. In contrast to the previous subsection, the investigated pretraining strategies were really heterogeneous, with at least one work for each category (predictive, generative, and contrastive) selected during the survey. Kyasseh *et al.* [84] proposed CLOCS, a family of patient-specific contrastive learning methods that they showed were able to outperform other baseline contrastive learning techniques, thus becoming the comparison element for other works. In particular, they presented three different approaches: the first, Contrastive Multi-Segment Coding (CMSC), exploits the temporal invariances in the ECG; the second, Contrastive Multi-Lead Coding (CMLC), exploits the spatial invariances in the ECG; and the latter, Contrastive Multi-Segment Multi-lead Coding (CMSMLC) combines the two previ-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

TABLE 1. Self-Supervised Learning on ECG

#	Author	Year	Upstream Task	Downstream Task	Dataset
1	Lee *et al.* [45]	2021	Predictive	Arrhythmia detection	CPSC2018 [79], PTB-XL [22], Chapman [80]
2	Luo *et al.* [81]	2021	Predictive	Arrhythmia classification	CinC 2017 [82]
3	Chen *et al.* [83]	2021	Contrastive	Arrhythmia classification	PTB-XL, CPSC2018, CinC 2017
4	Kiyasseh *et al.* [84]	2021	Contrastive (CLOCS)	Arrhythmia classification	CinC 2018 [85], CinC 2020 [86], Chapman, Cardiology [87]
5	Wei *et al.* [88]	2022	Contrastive	Arrhythmia classification	MIT-BIH [89], Chapman, private
6	Lan *et al.* [90]	2022	Contrastive (ISL)	Arrhythmia classification	Chapman, CPSC2018, PTB-XL
7	Phan *et al.* [91]	2022	Contrastive	Arrhythmia classification	CinC 2020
8	Grabowski *et al.* [92]	2022	Masking	Arrhythmia multiple	MIT-BIH
9	Oh *et al.* [93]	2022	Other	Arrhythmia classification	CinC 2021 [94], PTB-XL
10	Zhang *et al.* [95]	2023	Predictive	Atrial fibrillation detection	CinC 2017
11	Sarkar *et al.* [96]	2020	Predictive (multi-task)	Emotion recognition	SWELL [97], AMIGOS [98]
12	Sarkar *et al.* [99]	2021	Predictive	Stress multiple	FELICITy, AMIGOS, DREAMER [100], SWELL, WESAD [101]
13	Vazquez-Rodriguez *et al.* [102]	2022	Masking	Emotion recognition	ASCERTAIN [103], DREAMER, AMIGOS, PsPM[a]
14	Rabbani *et al.* [104]	2022	Contrastive	Stress classification	WESAD, RML
15	Gedon *et al.* [105]	2021	Masking	Pathology classification	CPSC2018, CODE [106], PTB-XL
16	Gopal *et al.* [107]	2021	Contrastive (3KG)	Pathology classification	CinC 2020
17	Lee *et al.* [108]	2021	Contrastive (VIbCReg)	Pathology classification	PTB-XL
18	Mehari *et al.* [109]	2022	Contrastive	Pathology classification	CinC 2020, CODE, Chapman, PTB-XL
19	Nakamoto *et al.* [110]	2022	Contrastive	Systolic dysfunction	UTokyo Private Dataset
20	Soltanieh *et al.* [111]	2022	Contrastive (SimCLR)	Pathology classification	PTB-XL
21	Yang *et al.* [112]	2022	Masking	Pathology classification	PTB-XL, CPSC2018, Chapman
22	Liu *et al.* [113]	2023	Contrastive	Pathology classification	Ningbo [114], PTB-XL, CPSC2018
23	Lai *et al.* [115]	2023	Contrastive	Pathology classification	Private

[a] [Online]. Available: https://bachlab.github.io/PsPM/opendata/

ous methods. Oh *et al.* [93] combined Kyasseh's CMSC with a transformer based self-supervised method (based on Wav2Vec 2.0 [116]), achieving performances superior to CLOCS. Moreover, they added a random lead masking module, which improves the model's robustness in the case of downstream tasks that accept an arbitrary number of ECG-leads. CLOCS was also considered as a comparison element by Chen *et al.* [83], which exported MoCo V2 to the ECG domain and used a combination of wavelet transform and random crop to generate the positive and negative pairs for the pretraining. Ultimately, Phan *et al.* [91] combined representations coming from both time and time-frequency modalities managed by two different backbone encoders pretrained with DINO [73]. In contrast with previous works, Lan *et al.* [90] designed an Intra- Inter-subject Self-supervised Learning (ISL) method for multivariate cardiac signals that tries to learn good representations of the ECG signal by learning distinct representations both at the heartbeat level (intra-subject) and at the subject level (inter-subject). The last selected contrastive learning proposal for arrhythmia classification is that of Wei *et al.* [88] with their "Contrastive HeartBeat", a novel method designed to learn patient-specific representations at the heartbeat level by considering as positive pairs all heartbeats of the same subject and as negative the others.

Again, contrastive learning was not the only approach investigated. Grabowski *et al.* [92] tested masked modeling on both classification and regression downstream problems. Furthermore, Zhang *et al.* [95] applied spatial and temporal signal manipulation to generate pseudo-labels for their predictive pretext task (transformation prediction). A predictive pretask was also chosen by Lee *et al.* [45] and Luo *et al.* [81] for arrhythmia detection and classification. The first constructed a pretraining based on the prediction of specific critical features extracted from the heartbeat with ECG delineation algorithms, while the second pretrained the model to assess if randomly selected pairs of ECG segments were adjacent or not.

As in the previous subsection, the large amount of supervision that can be provided from some of the available free open datasets resulted in model performances that were not always superior to their fully supervised counterparts. For example, Kyasseh *et al.* [84] reported an absolute drop in AUC ranging from 0.02 to 0.04 on different fine-tuning datasets, while Chen *et al.* [83] achieved an AUC improvement of 0.03 with a similar experimental setting. However, as reported in [84], [90], self-supervised learning was able to pro-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

duce similar results even when the fraction of labeled data used for fine-tuning was halved, thus mitigating the performance drop compared to other strategies. Ultimately, it is worth noting that it is possible to identify a positive progression in the achieved results (see the use of CLOCS in many other works as a baseline comparison), guided by the proposal of novel methods being able to include both physiological and subject-related information during pretraining.

3) emotion classification

Stress detection and emotion classification (e.g., prediction of the affective score) were the other two investigated downstream tasks. In particular, stress detection was studied by Rabbani *et al.* [104] and Sarkar *et al.* [99]. The first adopted the baseline SimCLR, while the second tried to assess maternal and fetal stress during pregnancy using a predictive multitask pretraining (e.g., prediction of different signal transformations). The same author extended this method to the emotion recognition problem [96], which was also studied by Rodriguez *et al.* [102], who chose masking modeling as the pretraining strategy.

Overall, selected works achieved performances superior to their fully supervised baseline. For example, Rodriguez *et al.* [102] achieved a mean absolute improvement of 0.03 over both accuracy and F1-score calculated on the AMIGOS dataset [98]. However, given the differences in the datasets used for the evaluation, it is impossible to compare results overall in order to extract a possible hierarchy in the pretraining strategy.

B. SELF-SUPERVISED LEARNING ON EEG

EEG is the other major type of biosignal where self-supervised learning has been applied. Here, SSL was employed for different downstream tasks such as sleep staging, seizure analysis, emotion classification, and motor imagery classification. The high number of downstream tasks, which were investigated using datasets provided by medical facilities, large clinical studies, or specific competitions, demonstrate how self-supervised learning could impact many real-world applications. For example, SSL-based sleep analysis can promote the development of novel deep learning-based automatic sleep scoring algorithms, which can eliminate some drawbacks of manual protocols [117], while SSL-based seizure analysis can improve the performance of automated detection systems, which allow an objective assessment of seizure frequency and a treatment tailored to the individual patient [118].

Differently from the results presented on ECG data, the choice of the pretext task depends on the study objective, with contrastive learning being slightly preferred overall. However, despite the chosen pretext task, works often attempt to include domain knowledge information about the EEG signal during pretraining, for example by considering the importance of frequency-based EEG analysis [119] or the similarity between resting state brain hemisphere activity [120].

1) sleep staging

Sleep staging, i.e., the problem of determining the patients' status (wake, light sleep, deep sleep, REM) during their sleep, was highly investigated, with 6 out of 23 EEG works selected during the survey. It is interesting to note that contrastive learning is predominant, with all works employing it to pretrain their models. Ren *et al.* [121] applied a modified version of contrastive predictive coding, while Jiang *et al.* [122] chose SimCLR. Yang *et al.* [123] proposed ContraWR, a novel approach that aims at solving the problem of negative sampling by using the average representation over the dataset as the only contrastive information. Another novel approach called SleepDPC was presented in Xiao *et al.* [124]. SleepDPC combines two different learning objectives during the pretraining: one, called predictive contrastive learning, uses the CPC-based Dense Predictive Coding (DPC) [125] as a reference; the other, called discriminative contrastive learning, tries to discern between temporary nearer or farther portions of the signal. Dense Predictive Coding was also part of the CoSleep method described in Ye *et al.* [119], which exploits multiple views of the EEG signal. In particular, DPC was first used to train from scratch two encoders, one for the time view and the other for the frequency view; then, contrastive multiview [126] was used to refine the weights of the two encoders. Finally, Lee *et al.* [127] presented SSLAPP, a hybrid approach based on the combination of a GAN-based generative pretext task and contrastive learning, achieving performances superior to CoSleep, SleepDPC, and other fully supervised strategies.

Based on the results presented on the SleepEDF dataset [128] which, as can be seen in 2, was used for the evaluation of all the proposed methods, most of the works achieved an accuracy and F1-score superior to fully supervised baselines, with only CoSleep and SleepDPC being left behind. For example, SSLAPP reported an absolute improvement on the F1-score of 0.03 but, more importantly, the achievement of similar results using only 10% of the labeled data. However, despite the overall improvement, no real superiority can be found among the various selected strategies.

2) seizure analysis

Unlike sleep staging, where contrastive learning was the primary choice, studies dealing with seizure analysis usually adopted predictive pretext tasks. Xu *et al.* [129] generated the pseudo-labels by applying a set of scaling transformations to only the EEGs of healthy subjects and pretrained the model to detect them. Tang *et al.* [130] use forecasting (future 12 seconds on a clip of 12 or 60 seconds) as the predictive pretext task, combining

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

SSL and graph neural networks for seizure analysis (detection and classification) for the first time. Das *et al.* [53] pretrained the model to reconstruct the original signal by its own corrupted version using different modification protocols, including masked modeling, thus exploiting a generative pretext task. Finally, Yang *et al.* [131] combined self-supervision with online learning and weak-supervision for patient-specific seizure forecasting.

Seizure analysis is really heterogeneous in the investigated types of learning problems and performance achieved. When it comes to seizure detection, for example, all the proposed methods were able to surpass fully supervised baselines. On the contrary, seizure classification (identification of seizure type) and forecasting pose more challenges. Hopefully, advancements in the research will reveal the potentiality of SSL for those problems as well.

3) motor imagery

Thanks to the BCI Competition IV [157], self-supervised learning was also extended to motor imagery, i.e., the mental execution of a movement without any overt movement or without any peripheral (muscle) activation [158]. Out of the three selected works, two used predictive pretext tasks. In particular, He *et al.* [147] pretrained their model to forecast a slice of the EEG signal given a set of past ones, while Ou *et al.* [149] randomly shuffled portions of the EEG signal and defined a binary classification task (signal segments in order or not). In contrast, Lotey *et al.* [145] assessed the impact of contrastive learning for cross-session motor imagery using the baseline SimCLR. However, they achieved a lower overall accuracy on the BCI Dataset 2a [150] compared to the forecasting proposal in He's work.

BCI competition datasets remain an important source of open datasets in the relative domain. They offer a common place to share and compare results achieved with different learning strategies, facilitating the advance of the research in this prominent field. Although SSL strategies were not able to achieve state-of-the-art results, which are still based on fully supervised methods [159], it is likely that their role will increase in future years, especially when the pretraining will be performed on multiple datasets to enhance the quality of the representations.

4) emotion recognition

Works employing self-supervised learning for emotion recognition varied in the choice of the pretext task. Xie *et al.* [140] applied six different transformations to EEG data and pretrained a multi-branch neural network to predict them. Zhang *et al.* [54] proposed GANSER, a generative self-supervised framework based on adversarial training. In particular, adversarial training is promoted through a masking operation and regulated by an augmentation factor designed to restrict the feature distribution difference between real EEG samples and the generated ones. Finally, Shen *et al.* [143] and Kan *et al.* [141] proposed two novel contrastive learning approaches. The first, Contrastive Learning for Inter-Subject Alignment (CLISA), tries to maximize the similarity in EEG signal representations across subjects who received the same emotional stimuli, hence without resorting to standard data augmentation procedures. The other, Group Meiosis Contrastive Learning (SGMC), adopted a genetically inspired data augmentation technique where positive and negative pairs are generated by grouping EEG samples sharing the same stimuli and then cross-exchanging (mixing) parts of their signal.

Overall, results on the widely used SEED [133] and DEAP [142] open datasets were superior to their fully supervised counterparts but comparable with each other. However, although minimal, it is worth noting that Xie's predictive pretext and Zhang's GANSER achieved state-of-the-art performances in the SEED and DEAP, respectively. This aspect highlights how, in emotion recognition, there is no clear superiority of one pretraining strategy over the others.

5) other or multiple classification tasks

Six other works adopted self-supervised learning on other downstream tasks or simply provided results on multiple applications. Mohsenvald *et al.* [132] provide an extensive analysis of the SimCLR contrastive learning framework on several downstream tasks. In particular, their analysis of the influence of the EEG sequence length, the applied data augmentation and the number of latent dimensions, as well as the role of the aggregation of heterogeneous datasets, is of great interest and provides a good insights on those aspects that are crucial for the efficient development of SSL strategies in the EEG domain. Banville *et al.* [135] evaluated three different pretext tasks (CPC and two predictive) on sleep staging and pathology classification. Wagh *et al.* [120] highlighted the importance of exploiting domain knowledge information from the EEG signal during pretraining and proposed an SSL method based on the combination of three different domain-guided pretext tasks (hemispheric symmetry, behavioral state estimation, and age contrastive). Zheng *et al.* [44] investigated the efficacy of SSL for anomaly detection on EEG data by designing a predictive pretask (3-class classification) where pseudo-labels were generated by locally increasing/decreasing the amplitude of the signal in the time domain or specific components in the frequency domain. Instead, Kostas *et al.* [136] designed BENDR, a novel method that combines a transformer-based framework with contrastive learning. Ultimately, Zygierewicz *et al.* [151] applied MoCo to memory-related neurofeedback data with the goal of identifying brain regions and frequency bands consistent with current neurophysiological knowledge of the processes critical to attention and

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

TABLE 2. Self-Supervised Learning works on EEG signal

#	Author	Year	Upstream Task	Downstream Task	Dataset
1	Mohsenvald *et al.* [132]	2020	Contrastive	Multiple	SEED [133], TUH [134], Sleep-EDF [128]
2	Banville *et al.* [135]	2021	Multiple	Multiple	CinC 2018 [85], TUH [134]
3	Kostas *et al.* [136]	2021	Other (BENDR)	Multiple	TUH
4	Wagh *et al.* [120]	2021	Other (CL + predictive)	Multiple	TUH, MPI LEMON [137]
5	Zheng *et al.* [44]	2022	Predictive	Anomaly detection	Private, CHB-MIT [138], UPMC [139]
6	Xie *et al.* [140]	2021	Predictive	Emotion classification	SEED
7	Kan *et al.* [141]	2022	Contrastive (SGMC)	Emotion classification	SEED, DEAP [142]
8	Shen *et al.* [143]	2022	Contrastive (CLISA)	Emotion classification	THU-EP [144], SEED
9	Zhang *et al.* [54]	2022	Generative (GANSER)	Emotion classification	DEAP, DREAMER [100], SEED
10	Lotey *et al.* [145]	2022	Contrastive	Motor Imagery classification	BCI-IV-2A [146]
11	He *et al.* [147]	2022	Predictive (forecasting)	Motor Imagery classification	MI-2 [148], BCI-IV-2A
12	Ou *et al.* [149]	2022	Predictive	Motor Imagery classification	BCI-IV-2A, BCI-IV-2B [150]
13	Zygierewicz *et al.* [151]	2022	Contrastive	Neurofeedback (memory)	Private
14	Xu *et al.* [129]	2020	Predictive	Seizure detection	UPMC
15	Tang *et al.* [130]	2021	Predictive (forecasting)	Seizure multiple	TUH
16	Das *et al.* [53]	2022	Generative	Seizure detection	TUH
17	Yang *et al.* [131]	2022	Predictive (forecasting)	Seizure forecasting	TUH, EPILEPSIAE [152], private
18	Jiang *et al.* [122]	2021	Contrastive	Sleep staging	Sleep-EDF, DOD [153]
19	Yang *et al.* [123]	2021	Contrastive (ContraWR)	Sleep staging	Sleep-EDF, SHHS [154], MGH [155]
20	Ye *et al.* [119]	2021	Contrastive (CoSleep)	Sleep staging	Sleep-EDF, ISRUC [156]
21	Xiao *et al.* [124]	2021	Contrastive (SleepDPC)	Sleep staging	Sleep-EDF, ISRUC
22	Ren *et al.* [121]	2022	Contrastive	Sleep staging	Sleep-EDF
23	Lee *et al.* [127]	2022	Other (SSLAPP)	Sleep staging	Sleep-EDF, ISRUC

working memory.

C. SELF-SUPERVISED LEARNING ON OTHER TYPES OF BIOSIGNAL

This section presents self-supervised learning applications on other types of biosignals such as EMG, eye tracking and other sensor data. Given the low number of selected works and the various biosignals included, it is difficult to identify a trend in the choice of the pretext task (see Table 3).

However, it is worth noting that works presented here are no less important than others from the previous sections, as the analysis of the biosignals included in this subsection is essential for many real-world applications, from the development of myoelectric prostheses to the support of older people's daily lives.

EMG certainly deserves a proper category because of its wide range of applications. However, only two studies adopting self-supervised learning on such data were found. In particular, Liu *et al.* [160] use contrastive learning (NeuroPose) to predict finger joint angles for 3D hand pose estimation from wearable EMG sensor data (8-channel armband), achieving good performances and demonstrating robustness to natural variation in sensor mounting positions or changes in the wrist position. Wu *et al.* [161] designed a novel self-supervised learning approach (Neuro2vec) for neurophysiological data based on masking pretext task applied to both the spatiotemporal and the frequency domains. They tested their approach on classification and regression tasks using EEG data and the NinaPro dataset [24], which is one of the biggest collections of open source datasets with EMG data. In the NinaPro dataset 5, they were able to achieve an absolute improvement of 0.03 both in accuracy and F1-score on the investigated classification task and a relative drop of 10% in the Mean Square Error on the regression task.

Regarding other modalities, Saeed *et al.* [162] exported self-supervised learning on **accelerometer** data for human activity recognition [163], a promising assistive field that can support older people's daily lives. In their work, they designed a multitask predictive approach based on the recognition of eight different signal transformations.

Considering **eye tracking** data, Mengoudi *et al.* [164] presented a predictive pretext task for their study. In particular, they tried to classify subjects with dementia, transferring the features learned during the pretraining to a support vector machine majority voting scheme.

Ultimately, Ballas *et al.* [165] designed Listen2YourHeart, a contrastive learning approach for Heart Murmur detection based on the baseline method SimCLR using **Phonocardiography** (PCG) data.

Overall, the investigated works demonstrated that SSL can be successfully applied to other types of biosignals, even when the amount of data available is not extremely high.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

TABLE 3. Self-Supervised Learning works on other types of Biosignals

#	Author	Year	Data type	Upstream Task	Downstream Task	Dataset
1	Saeed *et al.* [162]	2019	Activity data	Predictive	Activity classification	HHAR [166], UniMib [167], UCI HAR [168], MobiAct [169], WISDM [170], MotionSense [171]
2	Mengoudi *et al.* [164]	2020	Eye-tracking	Predictive	Dementia classification	Private
3	Ballas *et al.* [165]	2022	PCG	Contrastive	murmur prediction	CinC 2016 [172], CinC 2022 [173]
4	Liu *et al.* [160]	2022	EMG	Contrastive	3D hand pose estimation	Private
5	Wu *et al.* [161]	2022	EEG, EMG	Generative	Multiple	Sleep-EDF [128], Bonn EEG [23], NinaPro [24]

D. MULTIMODAL SELF-SUPERVISED LEARNING WITH BIOSIGNALS

Multimodal self-supervised learning with biosignals is the final category presented in this section. The number of works in this context is still limited (see Table 4), which highlights how efficiently combining information from different types of data is a difficult task. The modalities mainly analysed with self-supervised learning include combinations of EEG, ECG, EMG, and other data coming from wearable devices. Differently from the trend of single-modality SSL, most of the works chose predictive pretext tasks instead of contrastive learning. Furthermore, multimodal data are often treated simultaneously via multichannel architectures, with each modality having its own encoder and representations combined only on the network head.

SSL applications that employ data from wearable devices for medical tasks are still limited in number (despite the fact that many studies have been released for more industrial applications). Spathis *et al.* [174] investigated health and lifestyle monitoring with multimodal wearable data, designing a particular pretext task whose goal was to assess the subject's heart rate from other wearable data. Deldari *et al.* [175] presented COCOA, a contrastive learning approach designed to learn quality representations from multisensor data by computing the cross-correlation between different data modalities and minimizing the similarity between irrelevant instances. Their approach was tested on several downstream tasks (e.g., emotion recognition, sleep staging, human activity recognition) combining different biosignals such as EEG, ECG, EMG, EOG, and activity data from wearable devices, achieving overall results always superior to fully supervised strategies (absolute accuracy improvements range from 0.03 to more than 0.1). Saeed *et al.* [176] presented "sense and learn", a novel framework designed to learn general-purpose representations from multisensor data produced by omnipresent sensing systems. In their work, they compared several pretext tasks on multiple downstream tasks such as activity recognition, sleep scoring, and stress detection.

Three of the identified works applied self-supervised learning strategies to Intensive Care Unit (ICU) data. In particular, Chen *et al.* [177] proposed a novel method for the prediction of adverse surgical events. To accomplish that, they combined a set of static (e.g., covariates) and dynamic (e.g., biosignals) variables, pretraining the backbone module of the latter with a forecasting predictive pretext task. Weatherhead *et al.* [191] improved the baseline contrastive learning method TNC [71], pretraining the model on high-time resolution ICU data and evaluating it on several tasks such as the prediction of 12-hour in-hospital mortality, circulatory failure, and cardiopulmonary arrest. Ultimately, Tipirneni *et al.* [193] pretrained the model with a forecasting pretext task.

Considering works employing other combinations of biosignals, Lemkhenter *et al.* [188] investigated self-supervised learning for sleep scoring with polysomnography data (collection of EEG, EOG, EMG, and ECG acquired during sleep), adopting a predictive pretask task built on top of the model-agnostic meta-learning framework [196]. The learning problem was to detect if a training sample came or not from PhaseSwap [43], an operator that takes two signals as input and then combines the amplitude of the first with the phase of the second.

Thiam *et al.* [182] were the only one to propose a generative pretask on multimodal data. Using a multimodal deep denoising convolutional auto-encoder, they tested the pretrained model for the pain intensity classification, achieving state-of-the-art performances on the BioVid Heat Pain Database [183].

The last two selected multimodal approaches combine biosignals with video recordings. Leveraging a combination of EEG and facial activity data extracted from video, Das *et al.* [185] trained an explainable AI model to predict upcoming speech stuttering, while Martini *et al.* [178] showed the potentiality of multimodal self-supervised learning by combining stereoencephalography (SEEG) and video data to forecast seizure events in drug resistant epileptic subjects.

Overall, the listed works achieved performance comparable or superior to fully supervised baselines. Moreover, works like [188] (sleep staging) show how some downstream tasks can be treated both with single and multimodal approaches. In this regard, multimodality seems to help extract complementary representations,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

TABLE 4. Multimodal Self-Supervised Learning with Biosignals

#	Author	Year	Upstream Task	Downstream Task	Dataset
1	Chen *et al.* [177]	2021	Predictive (forecasting)	Surgical adverse events	Private
2	Martini *et al.* [178]	2021	Predictive	Seizure Forecasting	Private
3	Saeed *et al.* [176]	2021	Multiple	Multiple	HHAR [166], MobiAct [169], MotionSense [171], UCI HAR [168], HAPT [179], Sleep-EDF [128], MIT Driver DB [180], WiFi CSI [181]
4	Spathis *et al.* [174]	2021	Predictive	Subject Health	Private
5	Thiam *et al.* [182]	2021	Generative	Pain Classification	BioVid heat pain [183], SenseEmotion [184]
6	Das *et al.* [185]	2022	Predictive	Stuttering prediction	Private
7	Deldari *et al.* [175]	2022	Contrastive (COCOA)	Multiple	UCI HAR [168], SLEEP-EDF, PAMAP2 [186], WESAD [101], Opportunity [187]
8	Lemkhenter *et al.* [188]	2022	Predictive (PhaseSwap)	Sleep Scoring	SLEEP-EDF, ISRUC, UCD [189], CAP [190]
9	Weatherhead *et al.* [191]	2022	Contrastive	Multiple	HiRID [192]
10	Tipirneni *et al.* [193]	2022	Predictive	In-hospital mortality	MIMIC-III [194], CinC 2012 [195]

enhancing the quality of representations compared to single-modality SSL strategies.

VII. DISCUSSION AND OPEN CHALLENGES

This section aims at answering important questions that may arise from the analysis of the selected works: when self-supervised learning might be preferred to a standard fully supervised strategy; how data aggregation can improve the model's robustness; what is the role of the fine-tuning phase; what is the best pretext task to choose; what is the role of data augmentation during the pretraining; and how multimodality can benefit from this paradigm. Although some of these topics can be presented in general terms, particular focus will be given to the analysis of special aspects to consider when applying existing SSL techniques (which have succeeded in other time-series analysis tasks) to a specific biosignal analysis task. To make the narrative clearer and easier to follow, each topic will be presented concisely in a separate subsection, providing examples from the previous listed works whenever possible.

A. SUPERVISED VS SELF-SUPERVISED LEARNING

Overall, the analysis of the selected works has shown that self-supervised learning may improve the performance of the trained model and mitigate overfitting in most of the listed downstream tasks. Hence, it seems likely that this strategy can be useful when performing deep learning-based biosignal analysis. However, one must be cautious and consider some important aspects that may guide the researcher towards the choice of the most suitable training strategy.

First, it is important to address the amount of supervision which can be provided for the downstream task. If the amount of labeled data is sufficiently high, it is unlikely that SSL will boost performances in a statistically significant manner, especially when pretraining and fine-tuning is performed on the same single repository. However, it is difficult to find such datasets in the domain of biosignals. Few exceptions worth to be mentioned are the Temple University Hospital (TUH) dataset for EEG or the Computing in Cardiology (CinC) datasets for ECG. In fact, works employing such datasets (for example [113], [135]) were not always able to improve their performances with respect to fully supervised baselines. However, although results can be comparable, SSL has proven to lead to a better generalization of the problem, as a drastic decrease in the amount of supervision is translated into only a slight drop in performance, the opposite of fully supervised methods.

The second aspect to consider is the amount of external data that can be exploited during pretraining. Self-supervised learning's main goal is to provide a way to learn general-purpose features by exploiting large amounts of unlabeled data. The more data that can be fed into the network during pretraining, the more robust the learned features will be, as they come from a larger and more heterogeneous parterre of data. This has the potential to boost the performance on the downstream task, as reported in many of the selected works. A more in-depth analysis regarding the role of data aggregation in self-supervision will be done in the next subsection.

B. THE POWER OF DATA AGGREGATION

Regardless of the specific type of biosignal or the investigated clinical task, self-supervised learning pretraining has demonstrated that it can reach state-of-the-art performances when more datasets are simultaneously employed. However, although SSL approaches facilitate the aggregation of multiple repositories not acquired in the same experimental setting, this procedure is still not a common practice in biosignal analysis. There are indeed many studies that combine more than one dataset during pretraining, but their number is generally

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

limited to two or three repositories, usually acquired for the same medical purpose. On the contrary, works like [132] demonstrated how data aggregation might improve model performance even when records came from completely different experimental settings.

Practical limitations like the inability to standardize multiple datasets in an automatic and easy way certainly play a key role in the hindrance of such practice. In fact, biosignals are not only really complex to interpret but also suffer from great variability, which may come from experimental settings, acquisition protocols, storage modalities, and intra- and inter-subject variability. For example, EEG preprocessing includes not only data imputation, resampling, and filtering as for any other biosignals, but also the re-referencing to a common (or average) channel, the alignment to a unique template, and the interpolation of missing channels. Manually performing all these steps is an extremely time-consuming and discouraging task. However, as of now, no tools are designed to simultaneously preprocess and align multiple datasets of the same modality. Therefore, it could be of great interest for the research community to develop novel tools that are able to both perform consistent preprocessing on multiple datasets and integrate their functionalities with preexisting ones, such as EEGlab [197] for EEG or ECG-kit [198] for ECG, allowing to aggregate heterogeneous datasets for SSL applications.

Moreover, although there is already evidence that the aggregation of multiple datasets can improve the accuracy of downstream models [188], [199], it could be useful to further investigate the effect of massive data aggregation during pretraining and how the quality of the general-purpose features learned is affected by that. It could also be of great interest to understand whether this strategy could be exploited in advancing the problem of domain adaptation [200], i.e., the problem of avoiding significant performance degradation due to changes in the marginal distribution of the feature space (domain shift), which still remains a critical aspect in the biomedical domain.

C. THE CHOICE OF THE FINE-TUNING DATASET

When defining a self-supervised experimental pipeline, it is important to not only select the right pretraining datasets to aggregate but also the fine-tuning one. Unfortunately, considering the way self-supervised learning strategies are usually presented, fine-tuning seems to often take a back seat. However, this phase is no less important than the pretraining one, since model evaluation will be based on the performance metrics estimated from the test set of the fine-tuning dataset. Moreover, choosing the right fine-tuning dataset is important not only for model evaluation, but also to promote results replicability and facilitate the comparison between different approaches.

Regarding results replicability, one should opt as much as possible for free open repositories, or at least ones accessible up to a filled-out request form. While the use of private datasets is certainly not forbidden, especially during pretraining, it is also true that the community could benefit more from the introduction of novel strategies tested with only open datasets. The use of open data can, in fact, make results not only reproducible but also more reliable since nothing is hidden from the reader. Moreover, it encourages the use of the same dataset as well as the production and release of tools designed to preprocess and split it in a standardized way, which is a crucial step in the creation of useful benchmarks.

Regarding the comparison between different approaches, while in other fields such as computer vision the research community has adopted well-defined protocols (e.g., use of datasets with predefined test sets, use of the same combination of data augmentations, use of standard model architectures) to promote fair and robust comparison between the proposed strategies, in the biosignal domain this aspect remains an open challenge. In fact, given a specific downstream task, several factors, such as the choice of different fine-tuning datasets, the use of a different splitting strategy (subject-, session- or trial-based), or the way performance variability was assessed (repeated fine-tuning, leave one subject out cross-validation, pretraining with different subsets of data), often make it impossible to compare the presented results. While the splitting strategy and the performance variability assessment can change according to the experimental study, the choice of the specific fine-tuning dataset can be at least uniformed based on the investigated downstream task. To help readers choose the right fine-tuning repository, the following list of datasets often used for different downstream tasks is provided:

- **PTB−XL**: A large open dataset comprised of 21799 clinical 12-lead ECG records of 10 seconds length from 18869 patients. Each ECG was assigned a diagnostic label based on the evaluation of expert cardiologists. The number of labels can vary depending on the chosen experimental setting. Data can be directly downloaded.
- **CinC 2017**: another ECG dataset released for the Computing in Cardiology 2017 challenge. It includes 8528 single-lead ECGs with various types of arrhythmias diagnosed by expert cardiologists. It can be used in studies focused on the diagnosis of arrhythmias. Data can be directly downloaded.
- **TUH**: the largest EEG repository to date. It includes EEG records from 10874 subjects recorded at a minimum of 250 Hz with a 24- to 36-channel system. The dataset was also divided into several subsets annotated for specific case studies (e.g., TUAB for normal/abnormal classification, TUEP for epilepsy). Data can be accessed only after filling

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

out a request form.

- **DEAP** : an EEG dataset for emotion studies. It comprises EEG records from 32 subjects, with the possibility to download already processed samples. Data can be accessed only after filling out a request form, which must come from researchers with a permanent position at an academic or research institute.
- **BCI competition** : a set of datasets released for the BCI Competition IV. Widely used datasets include the datasets 2a and 2b for motor imagery with EEG data. Data can be directly downloaded.
- **NinaPro** : a large multimodal database aimed at fostering machine learning research on human, robotic and prosthetic hands. It comprises 10 datasets with EMG and other kinematic or inertial data acquired from subjects with intact or amputated hands. Data can be directly downloaded.
- **MIMIC** : a large multimodal dataset that included multimodal recordings from ICU patients. The datasets often employed are the MIMIC-II and MIMIC-III datasets, which can be accessed only after filling out a request form.
- **WESAD** : a multimodal dataset for wearable stress and affect detection comprised of physiological and motion data recorded from 15 subjects. Data can be directly downloaded.
- **Sleep−EDF** : a widely used dataset comprised of 197 polysomnographic sleep recordings. Despite its multimodality nature, this dataset is often employed in single-modality EEG sleep studies (see table 2). Data can be directly downloaded.
- **CinC 2018** : Another large sleep staging dataset composed of various physiological signals (ECG, EEG, EOG, and EMG) recorded from 1985 subjects. Data can be directly downloaded.

D. THE CHOICE OF THE PRETEXT TASK

Looking at the pure numbers, contrastive learning was the most chosen pretext task, outnumbering the sum of works adopting other methodologies. This aspect certainly reflects not only the ability of contrastive learning pretext to learn better general-purpose representations from the data compared to other approaches but also its easiness of adaption to the medical domain. In fact, self-supervised contrastive learning baseline approaches are fairly easy to implement and have lots of alternatives that, although similar, can fit specific experimental needs. Moreover, they can also be easily modified without actually changing their core parts. Many of the presented works, rather than designing completely novel approaches, slightly change baseline methods in order to incorporate specific medical domain knowledge. For example, some works proposed more biologically inspired data augmentation techniques [141], while others focused on the way similarity between pairs is evaluated, for example by modifying the objective learning function or the structure of the siamese network [108]. Specific examples of the incorporation of medical domain knowledge during pretraining can be found in the surveyed works. In particular, authors in [120] have presented an EEG-based multitask pretraining strategy which takes into account both similarities and dissimilarities in the activity of the left and right brain hemispheres but also considers the known effect on the EEG dynamic of both the age and behavioural state of the subject. In addition, although not classified as a contrastive learning pretext task, the method presented in [45] represents another example of domain knowledge incorporation since it is based on the prediction of characteristic features automatically extracted from the ECG signal (with standard procedures) and typically used by cardiologists for diagnostic purposes.

Although contrastive learning seems to generally perform well, discarding other pretraining strategies can be counterproductive. For example, predictive pretext tasks can lead to better results on some downstream tasks if properly designed, like motor imagery classification [147]. Moreover, they are still largely employed in multimodal approaches, where finding effective ways to assess similarities and dissimilarities in representations of different modalities for contrastive approaches is still an open challenge. Masked modeling was also successfully applied for several downstream tasks, although its paradigm is less open to novel implementation. However, when combined with other SSL strategies, especially when transformer architectures are involved [93], it could improve the model's performance and robustness.

Each pretraining technique has its own peculiarity; hence, it is reasonable to assume that the quality of the representations will be affected as well. In this context, it could be more valuable to investigate "hybrid" approaches, which incorporate the qualities of different methods, rather than trying to assess the best strategy among the categories. The combination of multiple pretext tasks might lead to more robust features, as they can instill in the feature extractor knowledge learned from very different tasks. An example of such a strategy can be found in [127], where contrastive (SimSiam) and generative (GAN-based) pretext tasks were combined to improve the quality of the representations.

In conclusion, it is probably still too early to understand what the best SSL pretext task category is for the analysis of biosignals, especially considering that the field is evolving quickly and some methods (e.g., generative pretext tasks) have such a limited number of applications. Future works and advancements in this domain will have the possibility of revealing which directions will be more effective.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

E. THE ROLE OF DATA AUGMENTATION

Data augmentations play a central role in affecting the quality of the representations learned during pretraining. They guide the network during the general-purpose feature learning process, consequently influencing its performance on the target task after fine-tuning. This fact is true not only in contrastive pretext tasks, where data augmentation is an essential part of the general workflow, but also in generative (e.g., reconstructive) and predictive strategies. Therefore, particular attention must be given to the design of augmentation methods, as wrong choices could deeply degrade the model's performance. Considering the field of application, it is extremely important to consider both the physiological nature of the signal and the prior medical knowledge about the target clinical task.

As for the signal's physiological nature, a data augmentation must generate a new version of the same data that not only preserves its physiological information but also does not diverge too much from the original dataset distribution. For example, a common employed data augmentation is the addition of generated noise or artifacts. In the biosignal domain, there are many known physiological sources of artifacts that could be exploited, such as the line noise, the drift artifact caused by changes in the electrodes' impedance, or the ocular and muscle artifacts typical of EEG data.

As for the medical knowledge of the target clinical task, while it is true that pretext tasks can produce robust features without any knowledge about the subsequent clinical task, it is also true that indirectly including such information in the model could be beneficial, even at the cost of reaching a worse loss minima during pretraining. For example, if the medical literature has already identified specific patterns which can be exploited to distinguish between normal (healthy) and abnormal (pathological) signals, it is important and reasonable to design data augmentations that will force the network to focus on such aspects. This, for example, applies to variations of the PQRST complex in ECG analysis or variations of the signal spectrum in specific bands in EEG applications.

Another key point to assess is how data augmentations are combined. While a single data augmentation chosen at random from a wide list could be a good initial strategy, compositions of multiple transformations can increase the sample's heterogeneity and produce more complex patterns, enhancing the learning process. In fact, as reported in [59], the composition makes the pretext task harder, but the quality of the representations improves dramatically. In the same work, the authors proposed a good pipeline to systematically study the impact of data augmentation, which was also used in [122] on EEG data. The results of both works demonstrated the superiority of data augmentation composition. However, it is also important not to stack too many augmentations, as the new transformed sample will be too noisy; hence, the trade-off between task complexity and quality or representation will probably be lost. A good compromise could be to apply a sequence of 2 augmentations, preceded by another physiologically invariant transformation, designed to increase the amount of training samples without actually changing the biological information of the original data. An example of such augmentation could be the EEG re-referencing to another channel, as suggested in [18], or the signal polarity inversion.

Despite the central role of data augmentation, the literature still lacks an extensive analysis of its role in SSL-based biosignal analysis. Aside from the previously mentioned work on EEG data, a similar analysis on ECG data is provided in [111]. However, no extensive study about their composition was performed.

F. THE CHALLENGE OF MULTIMODALITY

In the biomedical domain, multimodal data are often complementary with each other, meaning that each type of data (e.g., signals, images, text reports) can be used to extract unique latent representations to allow a better understanding of a pathology, even at the subject level. However, the analysis of the methods presented in the selected works certainly reveals how difficult it is to exploit the SSL paradigm in a multimodal environment. Two main reasons can explain this difficulty: the limited availability of multimodal datasets acquired for a specific task, and the challenging problem of effectively combining different modalities during pretraining.

As for the availability of multimodal datasets, there is no doubt that their collection within a unique experimental setting is extremely hard and costly. However, as reported in section VI-D, a common adopted strategy is to train a specific feature extractor for each data modality. This allows to overcome the problem of data availability by performing a two-step pretraining strategy. In the first step, each encoder can be trained separately by aggregating several unimodal repositories; then, multimodal data should be used to simultaneously optimize and align representations of all the feature extractors.

While this strategy allows for lessening the needs of multimodal repositories, the second problem, which is how to effective combine multiple data types, remains open. As of now, predictive pretext is the most chosen approach, given its lower computational requirement and easiness of implementation. However, predictive pretexts rely on the concatenation of the different embeddings only at the network head level (usually discarded during the model transfer phase) without actually promoting the alignment of different modalities at the backbone level. On the contrary, contrastive learning could be the most suitable type of pretext task in this context, as it allows improving the agreement

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

between representations of different modalities by projecting them in a common latent space used to calculate the contrastive loss (see COCOA [175]). The alignment of different modalities in a single space could have great potential in knowledge discovery scenarios, for example, by connecting the aligned embeddings to a common ontology. It could also open new possibilities in deep phenotyping and precision medicine [201], [202].

One medical area that could benefit most from multimodal applications is neuroscience. Neuroscience is extremely multimodal, with biosignals like EEG or EMG collected together with different types of images (e.g., positron emission tomography, optical coherence tomography, structural and functional magnetic resonance imaging) and tabular data. However, limited effort has been made to align images, signals, and clinical data, a procedure that could greatly improve the study of different neurological disorders and the understanding of the mechanisms behind their onset and progression.

Another application which could benefit from the use of multimodal self-supervised strategies is the management of chronic diseases through multimodal wearable data. As previously stated in section I, the role of wearable devices is constantly growing, and nowadays, people affected by chronic diseases like diabetes, coronary heart disease, or chronic obstructive pulmonary disease can heavily rely on them [203]. However, while different wearable devices can facilitate the monitoring of several physiological information, the introduction of deep learning-based decision support systems that can exploit them in real-world scenarios is still hindered by the high sources of variability (e.g., subject variability, sensor variability) associated with such data. In this context, the ability of self-supervised learning to improve model generalizability, as reported in other surveyed works, could help solve this problem. However, further investigations need to be performed, as the number of SSL-based works in this area is still limited.

VIII. CONCLUSIONS

Self-supervised learning represents a relatively recent and extremely powerful resource in the context of deep learning and, more generally, machine learning applications to different data modalities. In particular, the potential impact of self-supervised learning in biomedical sciences, where it's difficult to get large amounts of annotated data, is extremely high. While previous works reviewed SSL applications to biomedical images, this is the first review paper targeting SSL applications for the analysis of biosignals. The survey highlights how self-supervised learning has been widely adopted for various types of biosignals, including multimodal approaches. It also highlights how, despite its relatively young age, SSL can potentially solve the problem of learning robust representations from biosignals in situations where there is a limited amount of labeled data. However, several factors remain unclear and require further investigations, such as the choice of the pretext task, the data aggregation procedure, and the exploitation of biological information from biosignals during the pretraining phase. Despite these limitations, self-supervised learning has opened the path to a more robust and performant deep learning, which could finally bridge the gap between research and clinical applications. It also has the potential to make applications of deep learning in the biomedical domain (where it's more difficult to get data and annotations by experts) more substantial and to help face some open challenges (e.g., accountability, distribution shifts, robustness), which still hinder the reliability of AI for healthcare [204], [205].

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.
[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, Ieee, 2009.
[3] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al., "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," arXiv preprint arXiv:1712.01815, 2017.
[4] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., "Highly accurate protein structure prediction with alphafold," Nature, vol. 596, no. 7873, pp. 583–589, 2021.
[5] Z. Lu, "Pubmed and beyond: a survey of web tools for searching biomedical literature," Database, vol. 2011, 2011.
[6] V. Kaul, S. Enslin, and S. A. Gross, "History of artificial intelligence in medicine," Gastrointestinal endoscopy, vol. 92, no. 4, pp. 807–812, 2020.
[7] J. Egger, C. Gsaxner, A. Pepe, K. L. Pomykala, F. Jonske, M. Kurz, J. Li, and J. Kleesiek, "Medical deep learning—a systematic meta-review," Computer methods and programs in biomedicine, vol. 221, p. 106874, 2022.
[8] M. Wodzinski, T. Banzato, M. Atzori, V. Andrearczyk, Y. D. Cid, and H. Muller, "Training deep neural networks for small and highly heterogeneous mri datasets for cancer grading," in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1758–1761, IEEE, 2020.
[9] F. Wang, L. P. Casalino, and D. Khullar, "Deep learning in medicine—promise, progress, and challenges," JAMA internal medicine, vol. 179, no. 3, pp. 293–294, 2019.
[10] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," Journal of Big Data, vol. 6, no. 1, pp. 1–25, 2019.
[11] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "Ammu: a survey of transformer-based biomedical pretrained language models," Journal of biomedical informatics, p. 103982, 2021.
[12] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 11, pp. 4037–4058, 2020.
[13] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, et al., "Self-supervised speech representation learning: A review," IEEE Journal of Selected Topics in Signal Processing, 2022.
[14] G. Kahn, P. Abbeel, and S. Levine, "Badgr: An autonomous self-supervised learning-based navigation system," IEEE

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

Robotics and Automation Letters, vol. 6, no. 2, pp. 1312–1319, 2021.

[15] Z. Liu, A. Alavi, M. Li, and X. Zhang, "Self-supervised contrastive learning for medical time series: A systematic review," Sensors, vol. 23, no. 9, p. 4221, 2023.

[16] O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," Machine Learning with Applications, vol. 7, p. 100198, 2022.

[17] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, "3d self-supervised methods for medical imaging," Advances in neural information processing systems, vol. 33, pp. 18158–18172, 2020.

[18] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: A survey," PeerJ Computer Science, vol. 8, p. e1045, 2022.

[19] J. Xu, "A review of self-supervised learning methods in the field of medical image analysis," International Journal of Image, Graphics and Signal Processing (IJIGSP), vol. 13, no. 4, pp. 33–46, 2021.

[20] I. cheol Jeong, D. Bychkov, and P. C. Searson, "Wearable devices for precision medicine and health state monitoring," IEEE Transactions on Biomedical Engineering, vol. 66, no. 5, pp. 1242–1258, 2018.

[21] M. H. Rafiei, L. V. Gauthier, H. Adeli, and D. Takabi, "Self-supervised learning for electroencephalography," IEEE Transactions on Neural Networks and Learning Systems, 2022.

[22] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "Ptb-xl, a large publicly available electrocardiography dataset," Scientific data, vol. 7, no. 1, p. 154, 2020.

[23] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," Physical Review E, vol. 64, no. 6, p. 061907, 2001.

[24] M. Atzori, A. Gijsberts, S. Heynen, A.-G. M. Hager, O. Deriaz, P. Van Der Smagt, C. Castellini, B. Caputo, and H. Müller, "Building the ninapro database: A resource for the biorobotics community," in 2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), pp. 1258–1265, IEEE, 2012.

[25] D. Bansal, Real-Time Data Acquisition in Human Physiology: Real-Time Acquisition, Processing, and Interpretation—A MATLAB-Based Approach. Academic Press, 2021.

[26] L. Lu, J. Zhang, Y. Xie, F. Gao, S. Xu, X. Wu, Z. Ye, et al., "Wearable health devices in health care: narrative systematic review," JMIR mHealth and uHealth, vol. 8, no. 11, p. e18907, 2020.

[27] D. Rodbard, "Continuous glucose monitoring: a review of successes, challenges, and opportunities," Diabetes technology & therapeutics, vol. 18, no. S2, pp. S2–3, 2016.

[28] F. Kusumoto, ECG interpretation: from pathophysiology to clinical application. Springer Nature, 2020.

[29] E. Niedermeyer and F. L. da Silva, Electroencephalography: basic principles, clinical applications, and related fields. Lippincott Williams & Wilkins, 2005.

[30] M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A.-G. M. Hager, S. Elsig, G. Giatsidis, F. Bassetto, and H. Müller, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," Scientific data, vol. 1, no. 1, pp. 1–13, 2014.

[31] M. Cognolato, M. Atzori, and H. Müller, "Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances," Journal of rehabilitation and assistive technologies engineering, vol. 5, p. 2055668318773991, 2018.

[32] S. Yang, F. Zhu, X. Ling, Q. Liu, and P. Zhao, "Intelligent health care: Applications of deep learning in computational medicine," Frontiers in Genetics, p. 444, 2021.

[33] A. Yakimovich, A. Beaugnon, Y. Huang, and E. Ozkirimli, "Labels in a haystack: Approaches beyond supervised learning in biomedical applications," Patterns, vol. 2, no. 12, p. 100383, 2021.

[34] A. Chebli, A. Djebbar, and H. F. Marouani, "Semi-supervised learning for medical application: A survey," in 2018 International Conference on Applied Smart Systems (ICASS), pp. 1–9, IEEE, 2018.

[35] Z.-H. Zhou, "A brief introduction to weakly supervised learning," National science review, vol. 5, no. 1, pp. 44–53, 2018.

[36] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," IEEE Signal Processing Magazine, vol. 39, no. 3, pp. 42–62, 2022.

[37] T. S. Madhulatha, "An overview on clustering methods," arXiv preprint arXiv:1205.1117, 2012.

[38] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An overview of principal component analysis," Journal of Signal and Information Processing, vol. 4, no. 3B, p. 173, 2013.

[39] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150202, 2016.

[40] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," Proceedings of the IEEE, vol. 109, no. 1, pp. 43–76, 2020.

[41] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, "Self-supervised feature learning for 3d medical images by playing a rubik's cube," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22, pp. 420–428, Springer, 2019.

[42] S. Deldari, H. Xue, A. Saeed, J. He, D. V. Smith, and F. D. Salim, "Beyond just vision: A review on self-supervised representation learning on multimodal and temporal data," arXiv preprint arXiv:2206.02353, 2022.

[43] A. Lemkhenter and P. Favaro, "Boosting generalization in bio-signal classification by learning the phase-amplitude coupling," in Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28–October 1, 2020, Proceedings 42, pp. 72–85, Springer, 2021.

[44] Y. Zheng, Z. Liu, R. Mo, Z. Chen, W.-s. Zheng, and R. Wang, "Task-oriented self-supervised learning for anomaly detection in electroencephalography," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII, pp. 193–203, Springer, 2022.

[45] B. T. Lee, S. T. Kong, Y. Song, and Y. Lee, "Self-supervised learning with electrocardiogram delineation for arrhythmia detection," in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 591–594, IEEE, 2021.

[46] C. Zhang, C. Zhang, J. Song, J. S. K. Yi, K. Zhang, and I. S. Kweon, "A survey on masked autoencoder for self-supervised learning in vision and beyond," arXiv preprint arXiv:2208.00173, 2022.

[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[48] C. Zhang, C. Zhang, J. Song, J. S. K. Yi, K. Zhang, and I. S. Kweon, "A survey on masked autoencoder for self-supervised learning in vision and beyond," arXiv preprint arXiv:2208.00173, 2022.

[49] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," arXiv preprint arXiv:2003.05991, 2020.

[50] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," IEEE signal processing magazine, vol. 35, no. 1, pp. 53–65, 2018.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

[51] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009, 2022.

[52] W. Wang, Q. Tang, and K. Livescu, "Unsupervised pre-training of bidirectional speech encoders via masked reconstruction," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6889–6893, IEEE, 2020.

[53] S. Das, P. Pandey, and K. P. Miyapuram, "Improving self-supervised pretraining models for epileptic seizure detection from eeg data," arXiv preprint arXiv:2207.06911, 2022.

[54] Z. Zhang, S.-h. Zhong, and Y. Liu, "Ganser: A self-supervised data augmentation framework for eeg-based emotion recognition," IEEE Transactions on Affective Computing, 2022.

[55] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," Technologies, vol. 9, no. 1, p. 2, 2020.

[56] G. Koch, R. Zemel, R. Salakhutdinov, et al., "Siamese neural networks for one-shot image recognition," in ICML deep learning workshop, vol. 2, Lille, 2015.

[57] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv preprint arXiv:1807.03748, 2018.

[58] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 297–304, JMLR Workshop and Conference Proceedings, 2010.

[59] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in International conference on machine learning, pp. 1597–1607, PMLR, 2020.

[60] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738, 2020.

[61] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," arXiv preprint arXiv:2003.04297, 2020.

[62] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3733–3742, 2018.

[63] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., "Bootstrap your own latent-a new approach to self-supervised learning," Advances in neural information processing systems, vol. 33, pp. 21271–21284, 2020.

[64] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," arXiv preprint arXiv:2110.09348, 2021.

[65] X. Chen and K. He, "Exploring simple siamese representation learning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 15750–15758, 2021.

[66] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," Advances in neural information processing systems, vol. 33, pp. 9912–9924, 2020.

[67] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6707–6717, 2020.

[68] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in International Conference on Machine Learning, pp. 12310–12320, PMLR, 2021.

[69] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," arXiv preprint arXiv:2105.04906, 2021.

[70] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, "Whitening for self-supervised representation learning," in International Conference on Machine Learning, pp. 3015–3024, PMLR, 2021.

[71] S. Tonekaboni, D. Eytan, and A. Goldenberg, "Unsupervised representation learning for time series with temporal neighborhood coding," arXiv preprint arXiv:2106.00750, 2021.

[72] X. Chen*, S. Xie*, and K. He, "An empirical study of training self-supervised vision transformers," arXiv preprint arXiv:2104.02057, 2021.

[73] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660, 2021.

[74] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian, et al., "A cookbook of self-supervised learning," arXiv preprint arXiv:2304.12210, 2023.

[75] J. Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, and E. Azemi, "Subject-aware contrastive learning for biosignals," arXiv preprint arXiv:2007.04871, 2020.

[76] V. Gorade, A. Singh, and D. Mishra, "Large scale time-series representation learning via simultaneous low and high frequency feature bootstrapping," arXiv preprint arXiv:2204.11291, 2022.

[77] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik, "Self-supervised contrastive pre-training for time series via time-frequency consistency," arXiv preprint arXiv:2206.08496, 2022.

[78] K. Wickstrøm, M. Kampffmeyer, K. Ø. Mikalsen, and R. Jenssen, "Mixing up contrastive learning: Self-supervised representation learning for time series," Pattern Recognition Letters, vol. 155, pp. 54–61, 2022.

[79] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, et al., "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection," Journal of Medical Imaging and Health Informatics, vol. 8, no. 7, pp. 1368–1373, 2018.

[80] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," Scientific data, vol. 7, no. 1, p. 48, 2020.

[81] C. Luo, G. Wang, Z. Ding, H. Chen, and F. Yang, "Segment origin prediction: a self-supervised learning method for electrocardiogram arrhythmia classification," in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1132–1135, IEEE, 2021.

[82] G. D. Clifford, C. Liu, B. Moody, H. L. Li-wei, I. Silva, Q. Li, A. Johnson, and R. G. Mark, "Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017," in 2017 Computing in Cardiology (CinC), pp. 1–4, IEEE, 2017.

[83] H. Chen, G. Wang, G. Zhang, P. Zhang, and H. Yang, "Clecg: A novel contrastive learning framework for electrocardiogram arrhythmia classification," IEEE Signal Processing Letters, vol. 28, pp. 1993–1997, 2021.

[84] D. Kiyasseh, T. Zhu, and D. A. Clifton, "Clocs: Contrastive learning of cardiac signals across space, time, and patients," in International Conference on Machine Learning, pp. 5606–5615, PMLR, 2021.

[85] M. M. Ghassemi, B. E. Moody, L.-W. H. Lehman, C. Song, Q. Li, H. Sun, R. G. Mark, M. B. Westover, and G. D. Clifford, "You snooze, you win: the physionet/computing in cardiology challenge 2018," in 2018 Computing in Cardiology Conference (CinC), vol. 45, pp. 1–4, IEEE, 2018.

[86] E. A. P. Alday, A. Gu, A. J. Shah, C. Robichaux, A.-K. I. Wong, C. Liu, F. Liu, A. B. Rad, A. Elola, S. Seyedi, et al., "Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020," Physiological measurement, vol. 41, no. 12, p. 124003, 2020.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

[87] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nature medicine, vol. 25, no. 1, pp. 65–69, 2019.

[88] C. T. Wei, M.-E. Hsieh, C.-L. Liu, and V. S. Tseng, "Contrastive heartbeats: Contrastive learning for self-supervised ecg representation and phenotyping," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1126–1130, IEEE, 2022.

[89] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," IEEE engineering in medicine and biology magazine, vol. 20, no. 3, pp. 45–50, 2001.

[90] X. Lan, D. Ng, S. Hong, and M. Feng, "Intra-inter subject self-supervised learning for multivariate cardiac signals," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 4532–4540, 2022.

[91] T. Phan, D. Le, P. Brijesh, D. Adjeroh, J. Wu, M. O. Jensen, and N. Le, "Multimodality multi-lead ecg arrhythmia classification using self-supervised learning," in 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 01–04, IEEE, 2022.

[92] B. Grabowski, P. Głomb, W. Masarczyk, P. Pławiak, Ö. Yıldırım, U. R. Acharya, and R.-S. Tan, "Classification and self-supervised regression of arrhythmic ecg signals using convolutional neural networks," arXiv preprint arXiv:2210.14253, 2022.

[93] J. Oh, H. Chung, J.-m. Kwon, D.-g. Hong, and E. Choi, "Lead-agnostic self-supervised learning for local and global representations of electrocardiogram," in Conference on Health, Inference, and Learning, pp. 338–353, PMLR, 2022.

[94] M. A. Reyna, N. Sadr, E. A. P. Alday, A. Gu, A. J. Shah, C. Robichaux, A. B. Rad, A. Elola, S. Seyedi, S. Ansari, et al., "Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021," in 2021 Computing in Cardiology (CinC), vol. 48, pp. 1–4, IEEE, 2021.

[95] W. Zhang, S. Geng, and S. Hong, "A simple self-supervised ecg representation learning method via manipulated temporal–spatial reverse detection," Biomedical Signal Processing and Control, vol. 79, p. 104194, 2023.

[96] P. Sarkar and A. Etemad, "Self-supervised learning for ecg-based emotion recognition," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3217–3221, IEEE, 2020.

[97] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," in Proceedings of the 16th international conference on multimodal interaction, pp. 291–298, 2014.

[98] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," IEEE Transactions on Affective Computing, vol. 12, no. 2, pp. 479–493, 2018.

[99] P. Sarkar, S. Lobmaier, B. Fabre, D. González, A. Mueller, M. G. Frasch, M. C. Antonelli, and A. Etemad, "Detection of maternal and fetal stress from the electrocardiogram with self-supervised representation learning," Scientific reports, vol. 11, no. 1, pp. 1–10, 2021.

[100] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," IEEE journal of biomedical and health informatics, vol. 22, no. 1, pp. 98–107, 2017.

[101] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in Proceedings of the 20th ACM international conference on multimodal interaction, pp. 400–408, 2018.

[102] J. Vazquez-Rodriguez, G. Lefebvre, J. Cumin, and J. L. Crowley, "Transformer-based self-supervised learning for emotion recognition," in 2022 26th International Conference on Pattern Recognition (ICPR), pp. 2605–2612, IEEE, 2022.

[103] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," IEEE Transactions on Affective Computing, vol. 9, no. 2, pp. 147–160, 2016.

[104] S. Rabbani and N. Khan, "Contrastive self-supervised learning for stress detection from ecg data," Bioengineering, vol. 9, no. 8, p. 374, 2022.

[105] D. Gedon, A. H. Ribeiro, N. Wahlström, and T. B. Schön, "First steps towards self-supervised pretraining of the 12-lead ecg," in 2021 Computing in Cardiology (CinC), vol. 48, pp. 1–4, IEEE, 2021.

[106] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr, et al., "Automatic diagnosis of the 12-lead ecg using a deep neural network," Nature communications, vol. 11, no. 1, p. 1760, 2020.

[107] B. Gopal, R. Han, G. Raghupathi, A. Ng, G. Tison, and P. Rajpurkar, "3kg: contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations," in Machine Learning for Health, pp. 156–167, PMLR, 2021.

[108] D. Lee and E. Aune, "Computer vision self-supervised learning methods on time series," arXiv preprint arXiv:2109.00783, 2021.

[109] T. Mehari and N. Strodthoff, "Self-supervised representation learning from 12-lead ecg data," Computers in biology and medicine, vol. 141, p. 105114, 2022.

[110] M. Nakamoto, S. Kodera, H. Takeuchi, S. Sawano, S. Katsushika, K. Ninomiya, H. Akazawa, and I. Komuro, "Self-supervised contrastive learning for electrocardiograms to detect left ventricular systolic dysfunction," Proceedings of the Annual Conference of JSAI, vol. JSAI2022, 2022.

[111] S. Soltanieh, A. Etemad, and J. Hashemi, "Analysis of augmentations for contrastive ecg representation learning," in 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–10, IEEE, 2022.

[112] S. Yang, C. Lian, and Z. Zeng, "Masked autoencoder for ecg representation learning," in 2022 12th International Conference on Information Science and Technology (ICIST), pp. 95–98, IEEE, 2022.

[113] W. Liu, H. Zhang, S. Chang, H. Wang, J. He, and Q. Huang, "A joint cross-dimensional contrastive learning framework for 12-lead ecgs and its heterogeneous deployment on soc," Computers in Biology and Medicine, vol. 152, p. 106390, 2023.

[114] J. Zheng, H. Chu, D. Struppa, J. Zhang, S. M. Yacoub, H. El-Askary, A. Chang, L. Ehwerhemuepha, I. Abudayyeh, A. Barrett, et al., "Optimal multi-stage arrhythmia classification approach," Scientific reports, vol. 10, no. 1, p. 2898, 2020.

[115] J. Lai, H. Tan, J. Wang, L. Ji, J. Guo, B. Han, Y. Shi, Q. Feng, and W. Yang, "Practical intelligent diagnostic algorithm for wearable 12-lead ecg via self-supervised learning on large-scale dataset," Nature Communications, vol. 14, no. 1, p. 3741, 2023.

[116] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," Advances in neural information processing systems, vol. 33, pp. 12449–12460, 2020.

[117] H. Phan and K. Mikkelsen, "Automatic sleep staging of eeg signals: recent development, challenges, and future directions," Physiological Measurement, vol. 43, no. 4, p. 04TR01, 2022.

[118] A. Ulate-Campos, F. Coughlin, M. Gaínza-Lein, I. S. Fernández, P. Pearl, and T. Loddenkemper, "Automated seizure detection systems and their effectiveness for each type of seizure," Seizure, vol. 40, pp. 88–101, 2016.

[119] J. Ye, Q. Xiao, J. Wang, H. Zhang, J. Deng, and Y. Lin, "Cosleep: A multi-view representation learning framework for self-supervised learning of sleep stage classification," IEEE Signal Processing Letters, vol. 29, pp. 189–193, 2021.

[120] N. Wagh, J. Wei, S. Rawal, B. Berry, L. Barnard, B. Brinkmann, G. Worrell, D. Jones, and Y. Varatharajah, "Domain-guided self-supervision of eeg data improves down-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

stream classification performance and generalizability," in Machine Learning for Health, pp. 130–142, PMLR, 2021.

[121] C. Ren, L. Sun, and D. Peng, "A contrastive predictive coding-based classification framework for healthcare sensor data," Journal of Healthcare Engineering, vol. 2022, 2022.

[122] X. Jiang, J. Zhao, B. Du, and Z. Yuan, "Self-supervised contrastive learning for eeg-based sleep staging," in 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, IEEE, 2021.

[123] C. Yang, D. Xiao, M. B. Westover, and J. Sun, "Self-supervised eeg representation learning for automatic sleep staging," arXiv preprint arXiv:2110.15278, 2021.

[124] Q. Xiao, J. Wang, J. Ye, H. Zhang, Y. Bu, Y. Zhang, and H. Wu, "Self-supervised learning for sleep stage classification with predictive and discriminative contrastive coding," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1290–1294, IEEE, 2021.

[125] T. Han, W. Xie, and A. Zisserman, "Video representation learning by dense predictive coding," in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0, 2019.

[126] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, pp. 776–794, Springer, 2020.

[127] H. Lee, E. Seong, and D.-K. Chae, "Self-supervised learning with attention-based latent signal augmentation for sleep staging with limited labeled data," in Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22 (L. D. Raedt, ed.), pp. 3868–3876, International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[128] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg," IEEE Transactions on Biomedical Engineering, vol. 47, no. 9, pp. 1185–1194, 2000.

[129] J. Xu, Y. Zheng, Y. Mao, R. Wang, and W.-S. Zheng, "Anomaly detection on electroencephalography with self-supervised learning," in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 363–368, IEEE, 2020.

[130] S. Tang, J. A. Dunnmon, K. Saab, X. Zhang, Q. Huang, F. Dubost, D. L. Rubin, and C. Lee-Messer, "Self-supervised graph neural networks for improved electroencephalographic seizure analysis," arXiv preprint arXiv:2104.08336, 2021.

[131] Y. Yang, N. D. Truong, J. K. Eshraghian, A. Nikpour, and O. Kavehei, "Weak self-supervised learning for seizure forecasting: a feasibility study," Royal Society Open Science, vol. 9, no. 8, p. 220374, 2022.

[132] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive representation learning for electroencephalogram classification," in Machine Learning for Health, pp. 238–253, PMLR, 2020.

[133] W. Liu, W.-L. Zheng, Z. Li, S.-Y. Wu, L. Gan, and B.-L. Lu, "Identifying similarities and differences in emotion recognition with eeg and eye movements among chinese, german, and french people," Journal of Neural Engineering, vol. 19, no. 2, p. 026012, 2022.

[134] I. Obeid and J. Picone, "The temple university hospital eeg data corpus," Frontiers in neuroscience, vol. 10, p. 196, 2016.

[135] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort, "Uncovering the structure of clinical eeg signals with self-supervised learning," Journal of Neural Engineering, vol. 18, no. 4, p. 046020, 2021.

[136] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data," Frontiers in Human Neuroscience, vol. 15, p. 653659, 2021.

[137] A. Babayan, M. Erbey, D. Kumral, J. D. Reinelt, A. M. Reiter, J. Röbbig, H. L. Schaare, M. Uhlig, A. Anwander, P.-L. Bazin, et al., "A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral physiology in young and old adults," Scientific data, vol. 6, no. 1, pp. 1–21, 2019.

[138] A. H. Shoeb, Application of machine learning to epileptic seizure onset detection and treatment. PhD thesis, Massachusetts Institute of Technology, 2009.

[139] A. Temko, A. Sarkar, and G. Lightbody, "Detection of seizures in intracranial eeg: Upenn and mayo clinic's seizure detection challenge," in 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6582–6585, IEEE, 2015.

[140] Z. Xie, M. Zhou, and H. Sun, "A novel solution for eeg-based emotion recognition," in 2021 IEEE 21st International Conference on Communication Technology (ICCT), pp. 1134–1138, IEEE, 2021.

[141] H. Kan, J. Yu, J. Huang, Z. Liu, and H. Zhou, "Self-supervised group meiosis contrastive learning for eeg-based emotion recognition," arXiv preprint arXiv:2208.00877, 2022.

[142] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," IEEE transactions on affective computing, vol. 3, no. 1, pp. 18–31, 2011.

[143] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, "Contrastive learning of subject-invariant eeg representations for cross-subject emotion recognition," IEEE Transactions on Affective Computing, 2022.

[144] X. Hu, F. Wang, and D. Zhang, "Similar brains blend emotion in similar ways: Neural representations of individual difference in emotion profiles," Neuroimage, vol. 247, p. 118819, 2022.

[145] T. Lotey, P. Keserwani, G. Wasnik, and P. P. Roy, "Cross-session motor imagery eeg classification using self-supervised contrastive learning," in 2022 26th International Conference on Pattern Recognition (ICPR), pp. 975–981, IEEE, 2022.

[146] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "Bci competition 2008–graz data set a," Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology, vol. 16, pp. 1–6, 2008.

[147] Y. He, Z. Lu, J. Wang, S. Ying, and J. Shi, "A self-supervised learning based channel attention mlp-mixer network for motor imagery decoding," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 2406–2417, 2022.

[148] X. Ma, S. Qiu, and H. He, "Multi-channel eeg recording during motor imagery of different joints from the same limb," Scientific data, vol. 7, no. 1, p. 191, 2020.

[149] Y. Ou, S. Sun, H. Gan, R. Zhou, and Z. Yang, "An improved self-supervised learning for eeg classification," Math. Biosci. Eng, vol. 19, pp. 6907–6922, 2022.

[150] R. Leeb, C. Brunner, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "Bci competition 2008–graz data set b," Graz University of Technology, Austria, pp. 1–6, 2008.

[151] J. Żygierewicz, R. A. Janik, I. T. Podolak, A. Drozd, U. Malinowska, M. Poziomska, J. Wojciechowski, P. Ogniewski, P. Niedbalski, I. Terczynska, et al., "Decoding working memory-related information from repeated psychophysiological eeg experiments using convolutional and contrastive neural networks," Journal of Neural Engineering, vol. 19, no. 4, p. 046053, 2022.

[152] J. Klatt, H. Feldwisch-Drentrup, M. Ihle, V. Navarro, M. Neufang, C. Teixeira, C. Adam, M. Valderrama, C. Alvarado-Rojas, A. Witon, et al., "The epilepsiae database: An extensive electroencephalography database of epilepsy patients," 2012.

[153] A. Guillot, F. Sauvet, E. H. During, and V. Thorey, "Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging," IEEE transactions on neural systems and rehabilitation engineering, vol. 28, no. 9, pp. 1955–1965, 2020.

[154] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The national sleep research resource: towards a sleep data

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

commons," Journal of the American Medical Informatics Association, vol. 25, no. 10, pp. 1351–1358, 2018.

[155] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, and M. T. Bianchi, "Expert-level sleep scoring with deep neural networks," Journal of the American Medical Informatics Association, vol. 25, no. 12, pp. 1643–1650, 2018.

[156] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "Isruc-sleep: A comprehensive public dataset for sleep researchers," Computer methods and programs in biomedicine, vol. 124, pp. 180–192, 2016.

[157] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. Mueller-Putz, et al., "Review of the bci competition iv," Frontiers in neuroscience, p. 55, 2012.

[158] T. Mulder, "Motor imagery and action observation: cognitive tools for rehabilitation," Journal of neural transmission, vol. 114, pp. 1265–1278, 2007.

[159] H. Altaheri, G. Muhammad, and M. Alsulaiman, "Physics-informed attention temporal convolutional network for eeg-based motor imagery classification," IEEE Transactions on Industrial Informatics, vol. 19, no. 2, pp. 2249–2258, 2022.

[160] Y. Liu, S. Zhang, and M. Gowda, "A practical system for 3d hand pose tracking using emg wearables with applications to prosthetics and user interfaces," IEEE Internet of Things Journal, 2022.

[161] D. Wu, S. Li, J. Yang, and M. Sawan, "neuro2vec: Masked fourier spectrum prediction for neurophysiological representation learning," arXiv preprint arXiv:2204.12440, 2022.

[162] A. Saeed, T. Ozcelebi, and J. Lukkien, "Multi-task self-supervised learning for human activity detection," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 3, no. 2, pp. 1–30, 2019.

[163] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," IEEE communications surveys & tutorials, vol. 15, no. 3, pp. 1192–1209, 2012.

[164] K. Mengoudi, D. Ravi, K. X. Yong, S. Primativo, I. M. Pavisic, E. Brotherhood, K. Lu, J. M. Schott, S. J. Crutch, and D. C. Alexander, "Augmenting dementia cognitive assessment with instruction-less eye-tracking tests," IEEE journal of biomedical and health informatics, vol. 24, no. 11, pp. 3066–3075, 2020.

[165] A. Ballas, V. Papapanagiotou, A. Delopoulos, and C. Diou, "Listen2yourheart: A self-supervised approach for detecting murmur in heart-beat sounds," in 2022 Computing in Cardiology (CinC), vol. 498, pp. 1–4, IEEE, 2022.

[166] A. Stisen, H. Blunck, S. Bhattacharya, T. S. Prentow, M. B. Kjærgaard, A. Dey, T. Sonne, and M. M. Jensen, "Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition," in Proceedings of the 13th ACM conference on embedded networked sensor systems, pp. 127–140, 2015.

[167] D. Micucci, M. Mobilio, and P. Napoletano, "Unimib shar: A dataset for human activity recognition using acceleration data from smartphones," Applied Sciences, vol. 7, no. 10, p. 1101, 2017.

[168] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz, et al., "A public domain dataset for human activity recognition using smartphones.," in Esann, vol. 3, p. 3, 2013.

[169] C. Chatzaki, M. Pediaditis, G. Vavoulas, and M. Tsiknakis, "Human daily activity and fall recognition using a smartphone's acceleration sensor," in Information and Communication Technologies for Ageing Well and e-Health: Second International Conference, ICT4AWE 2016, Rome, Italy, April 21-22, 2016, Revised Selected Papers 2, pp. 100–118, Springer, 2017.

[170] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," ACM SigKDD Explorations Newsletter, vol. 12, no. 2, pp. 74–82, 2011.

[171] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Protecting sensory data against sensitive inferences," in Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems, pp. 1–6, 2018.

[172] G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark, "Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016," in 2016 Computing in cardiology conference (CinC), pp. 609–612, IEEE, 2016.

[173] M. A. Reyna, Y. Kiarashi, A. Elola, J. Oliveira, F. Renna, A. Gu, E. A. Perez-Alday, N. Sadr, A. Sharma, S. Mattos, et al., "Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022," medRxiv, pp. 2022–08, 2022.

[174] D. Spathis, I. Perez-Pozuelo, S. Brage, N. J. Wareham, and C. Mascolo, "Self-supervised transfer learning of physiological representations from free-living wearable data," in Proceedings of the Conference on Health, Inference, and Learning, pp. 69–78, 2021.

[175] S. Deldari, H. Xue, A. Saeed, D. V. Smith, and F. D. Salim, "Cocoa: Cross modality contrastive learning for sensor data," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 6, no. 3, pp. 1–28, 2022.

[176] A. Saeed, V. Ungureanu, and B. Gfeller, "Sense and learn: Self-supervision for omnipresent sensors," Machine Learning with Applications, vol. 6, p. 100152, 2021.

[177] H. Chen, S. M. Lundberg, G. Erion, J. H. Kim, and S.-I. Lee, "Forecasting adverse surgical events using self-supervised transfer learning for physiological signals," NPJ Digital Medicine, vol. 4, no. 1, p. 167, 2021.

[178] M. L. Martini, A. A. Valliani, C. Sun, A. B. Costa, S. Zhao, F. Panov, S. Ghatan, K. Rajan, and E. K. Oermann, "Deep anomaly detection of seizures with paired stereoelectroencephalography and video recordings," Scientific reports, vol. 11, no. 1, p. 7482, 2021.

[179] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," Neurocomputing, vol. 171, pp. 754–767, 2016.

[180] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," IEEE Transactions on intelligent transportation systems, vol. 6, no. 2, pp. 156–166, 2005.

[181] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using wifi channel state information," IEEE Communications Magazine, vol. 55, no. 10, pp. 98–104, 2017.

[182] P. Thiam, H. Hihn, D. A. Braun, H. A. Kestler, and F. Schwenker, "Multi-modal pain intensity assessment based on physiological signals: A deep learning perspective," Frontiers in Physiology, vol. 12, p. 720464, 2021.

[183] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A. O. Andrade, and G. M. da Silva, "The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system," in 2013 IEEE international conference on cybernetics (CYBCO), pp. 128–131, IEEE, 2013.

[184] M. Velana, S. Gruss, G. Layher, P. Thiam, Y. Zhang, D. Schork, V. Kessler, S. Meudt, H. Neumann, J. Kim, et al., "The senseemotion database: A multimodal database for the development and systematic validation of an automatic pain-and emotion-recognition system," in Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction: 4th IAPR TC 9 Workshop, MPRSS 2016, Cancun, Mexico, December 4, 2016, Revised Selected Papers 4, pp. 127–139, Springer, 2017.

[185] A. Das, J. Mock, F. Irani, Y. Huang, P. Najafirad, and E. Golob, "Multimodal explainable ai predicts upcoming speech behavior in adults who stutter," Frontiers in Neuroscience, p. 1200, 2022.

[186] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in 2012 16th international symposium on wearable computers, pp. 108–109, IEEE, 2012.

[187] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, et al., "Collecting complex activity datasets in highly rich networked sensor environments," in 2010 Seventh international conference on networked sensing systems (INSS), pp. 233–240, IEEE, 2010.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3344531

[188] A. Lemkhenter and P. Favaro, "Towards sleep scoring generalization through self-supervised meta-learning," in 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2961–2966, IEEE, 2022.
[189] C. Heneghan, "St. vincent's university hospital/university college dublin sleep apnea database," 2011.
[190] M. G. Terzano, L. Parrino, A. Sherieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky, A. Rosa, et al., "Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (cap) in human sleep," Sleep medicine, vol. 2, no. 6, pp. 537–554, 2001.
[191] A. Weatherhead, R. Greer, M.-A. Moga, M. Mazwi, D. Eytan, A. Goldenberg, and S. Tonekaboni, "Learning unsupervised representations for icu timeseries," in Conference on Health, Inference, and Learning, pp. 152–168, PMLR, 2022.
[192] M. Faltys, M. Zimmermann, X. Lyu, M. Hüser, S. Hyland, G. Rätsch, and T. Merz, "Hirid, a high time-resolution icu dataset (version 1.1. 1)," Physio. Net, vol. 10, 2021.
[193] S. Tipirneni and C. K. Reddy, "Self-supervised transformer for sparse and irregularly sampled multivariate clinical timeseries," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 16, no. 6, pp. 1–17, 2022.
[194] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," Scientific data, vol. 3, no. 1, pp. 1–9, 2016.
[195] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark, "Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012," in 2012 Computing in Cardiology, pp. 245–248, IEEE, 2012.
[196] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in International conference on machine learning, pp. 1126–1135, PMLR, 2017.
[197] C. Brunner, A. Delorme, and S. Makeig, "Eeglab–an open source matlab toolbox for electrophysiological research," Biomedical Engineering/Biomedizinische Technik, vol. 58, no. SI-1-Track-G, p. 000010151520134182, 2013.
[198] A. Demski and M. L. Soria, "Ecg-kit: a matlab toolbox for cardiovascular signal processing," Journal of open research software, vol. 4, no. 1, 2016.
[199] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive representation learning for electroencephalogram classification," in Machine Learning for Health, pp. 238–253, PMLR, 2020.
[200] H. Guan and M. Liu, "Domain adaptation for medical image analysis: a survey," IEEE Transactions on Biomedical Engineering, vol. 69, no. 3, pp. 1173–1185, 2021.
[201] P. N. Robinson, "Deep phenotyping for precision medicine," Human mutation, vol. 33, no. 5, pp. 777–780, 2012.
[202] N. Marini, S. Marchesin, S. Otálora, M. Wodzinski, A. Caputo, M. Van Rijthoven, W. Aswolinskiy, J.-M. Bokhorst, D. Podareanu, E. Petters, et al., "Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations," NPJ digital medicine, vol. 5, no. 1, p. 102, 2022.
[203] Y. Guo, X. Liu, S. Peng, X. Jiang, K. Xu, C. Chen, Z. Wang, C. Dai, and W. Chen, "A review of wearable and unobtrusive sensing technologies for chronic disease management," Computers in Biology and Medicine, vol. 129, p. 104163, 2021.
[204] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg, "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai," in Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 624–635, 2021.
[205] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," IEEE Reviews in Biomedical Engineering, vol. 14, pp. 156–180, 2020.

FEDERICO DEL PUP received a B.Sc. and a M.Sc. in Bioengineering from the University of Padua, Italy, in 2019 and 2022. He is currently pursuing the Ph.D. degree in Bioengineering at the Department of Information Engineering of the University of Padua, Italy. Since 2022, He has also been a researcher at the Department of Neuroscience of the University of Padua. His research interests include the analysis of unsupervised Machine/Deep Learning techniques for medical data analytic and knowledge discovery, with particular focus on multimodal approaches.

MANFREDO ATZORI received a M.Sc. in Physics and a Ph.D. in Bioengineering in 2006 and 2009 from the University of Padova, Italy. He is research scientist at the Institute of Information Systems of the University of Applied Sciences Western Switzerland (HES-SO Valais), Assistant Professor at the Department of Neuroscience of the University of Padova and he is the Scientific Coordinator of the Horizon 2020 project ExaMode, targeting multimodal weakly-supervised knowledge discovery in digital pathology. He has also been the coordinator of the Hasler Fundation financed ProHand project, targeting the development of 3D printed robotic prosthetic hands controlled via machine learning approaches. Between 2016 and 2019, Dr. Atzori had a leading role in the MeganePro Project, which aimed at improving robotic prosthesis control with eye-hand coordination and at better understanding the neurocognitive effects of amputations.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/